THE UNIVERSITY
of EDINBURGH

A Semi-automated Framework for the Analytical Use of Gene-centric Data with Biological Ontologies

Xin He

Doctor of Philosophy

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh

2017

Abstract

Motivation Translational bioinformatics(TBI) has been defined as 'the development and application of informatics methods that connect molecular entities to clinical entities' [1], which has emerged as a systems theory approach to bridge the huge wealth of biomedical data into clinical actions using a combination of innovations and resources across the entire spectrum of biomedical informatics approaches [2]. The challenge for TBI is the availability of both comprehensive knowledge based on genes and the corresponding tools that allow their analysis and exploitation.

Traditionally, biological researchers usually study one or only a few genes at a time, but in recent years high throughput technologies such as gene expression microarrays, protein mass-spectrometry and next-generation DNA and RNA sequencing have emerged that allow the simultaneous measurement of changes on a genome-wide scale. These technologies usually result in large lists of interesting genes, but meaningful biological interpretation remains a major challenge. Over the last decade, enrichment analysis has become standard practice in the analysis of such gene lists, enabling systematic assessment of the likelihood of differential representation of defined groups of genes compared to suitably annotated background knowledge. The success of such analyses are highly dependent on the availability and quality of the gene annotation data.

For many years, genes were annotated by different experts using inconsistent, non-standard terminologies. Large amounts of variation and duplication in these unstructured annotation sets, made them unsuitable for principled quantitative analysis. More recently, a lot of effort has been put into the development and use of structured, domain specific vocabularies to annotate genes. The Gene Ontology is one of the most successful examples of this where genes are annotated with terms from three main clades; biological process, molecular function and cellular component. However, there are many other established and emerging ontologies to aid biological data interpretation, but are rarely used. For the same reason, many bioinformatic tools only support analysis analysis using the Gene Ontology.

The lack of annotation coverage and the support for them in existing analytical tools to aid biological interpretation of data has become a major limitation to their utility and uptake. Thus, automatic approaches are needed to facilitate the transformation of unstructured data to unlock the potential of all ontologies, with corresponding bioinformatics tools to support their interpretation.

Approaches In this thesis, firstly, similar to the approach in [3,4], I propose a series of computational approaches implemented in a new tool *OntoSuite-Miner* to address the ontology based gene association data integration challenge. This approach uses NLP based text mining methods for ontology based biomedical text mining. What differentiates my approach from other approaches is that I integrate two of the most wildly used NLP modules into the framework, not only increasing the confidence of the text mining results, but also providing an annotation score for each mapping, based on the number of pieces of evidence in the literature and the number of NLP modules that agreed with the mapping. Since heterogeneous data is important in understanding human disease, the approach was designed to be generic, thus the ontology based annotation generation can be applied to different sources and can be repeated with different ontologies. Secondly, in respect of the second challenge proposed by TBI, to increase the statistical power of the annotation enrichment analysis, I propose *OntoSuite-Analytics*, which integrates a collection of enrichment analysis methods into a unified open-source software package named *topOnto*, in the statistical programming language R. The package supports enrichment analysis across multiple ontologies with a set of implemented statistical/topological algorithms, allowing the comparison of enrichment results across multiple ontologies and between different algorithms.

Results The methodologies described above were implemented and a Human Disease Ontology (HDO) based gene annotation database was generated by mining three publicly available database, OMIM, GeneRIF and Ensembl variation. With the availability of the HDO annotation and the corresponding ontology enrichment analysis tools in *topOnto*, I profiled 277 gene classes with human diseases and generated 'disease environments' for 1310 human diseases. The exploration of the disease profiles and disease environment provides an overview of known disease knowledge and provides new insights into disease mechanisms. The integration of multiple ontologies into a disease context demonstrates how 'orthogonal' ontologies can lead to biological insight that would have been missed by more traditional single ontology analysis.

Acknowledgements

I would like to thank many people who provided great help to not only my PhD study but the life in UK during the past years. They have made it possible for me to complete this thesis.

I am eternally grateful to my supervisor Ian, for providing me with the opportunity to undertake a PhD and for guidance during my PhD. You have put a considerable amount of time into meetings and discussions which were invaluable for the progress of my PhD. You have been incredibly patient and provided great encouragement when I met difficulties. This work could not have been completed without you. I also thank my second supervisor Douglas, for his guidance, critical advice and comments on my PhD work. I also appreciate the funding and working opportunities he provided. I also like to thank Colin for a lot insightful suggestions in my research. It is a great pleasure to work in such a friendly environment in the the Informatics Forum. Id like to thank all my friends and colleagues here.

I would like to express my special appreciation and thanks to my family. Words cannot express how grateful I am to my mother, father, my mother-in law and father-in-law for all of the sacrifices that they made on my behalf, and for the great love, encouragement, and confidence they have placed in me from start to finish. In particular, I would like express appreciation to my beloved wife LuLu, who was always supportive. Last but not least, I want to give special thanks to my 4 months old daughter Joyce, thank you for waking me up every night during the final stage of my PhD to complete this thesis.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Xin He*)

Table of Contents

List of Figures

List of Tables

Notation

Conventions

Italic text is used for gene symbols, software packages names and algorithm names. Except otherwise defined, *MeM* refers to the MetaMap program and *NcA* refers to the NCBO Annotator. '#' was used to represent the 'Number of'.

Abbreviation Reference

Abbreviation	Meaning
OMIM	Online Mendelian Inheritance in Man
GeneRIF	Gene Reference Into Function
HDGDB	Human Disease Gene Database
DAG	Directed Acyclic Graph
TPR	True Path Rule
SNP	Single Nucleotide Polymorphisms
GDA	Gene Disease Association
CS	Confidence Score

Ontology abbreviation	Meaning
CO	Chromosome Ontology
PCO	Panther Protein Class Ontology
RPO	Reactome Pathway Ontology
HDO	Human Disease Ontology
HPO	Human Phenotype Ontology
GOBP	Gene Ontology-Biological Process

Ontology abbreviation	Meaning
GOMF	Gene Ontology-Molecular Function
GOCC	Gene Ontology-Cellular Component
MeSH	Medical Subject Headings

Chapter 1

Introduction

The study of complex diseases requires the effective integration and analysis of disparate features that originate from genotypic, phenotypic, and environmental sources. Instead of a microscopic approach which focus on detailed analyses of a single data type, a macroscopic approach offers a holistic view for exploring complex diseases as systems by coalescing many heterogeneous data types. Translational bioinformatics(TBI), defined as "the development and application of informatics methods that connect molecular entities to clinical entities" [1], has thus emerged as a systems theory approach to bridge the huge wealth of biomedical data into clinical actions using a combination of innovations and resources across the entire spectrum of biomedical informatics approaches [2]. By the integrative exploitation of information, TBI will enable a deeper understanding of disease mechanisms and provide a new paradigm for the study and treatment of disease. The challenge is the availability of both comprehensive sources of gene annotation data and tools that allow their analysis and exploitation.

Traditionally, biological researchers usually study one gene or only a few genes at a time. Nowadays, new high-throughput scanning approaches such as DNA-Seq [5], RNA-Seq [6] and RNA microarrays [7] allow researchers to simultaneously measure the properties of genome-wide genes and proteins across entire genomes. These high-throughput technologies usually generate large gene lists, however, to understand the biological interpretation of these potentially interesting genes to gain disease insight is still a major challenge.

1.1 Existing gene annotation data

The first challenge is the lack of structured gene annotation data. Biomedical sciences are facing an enormous increase of data available from public sources. However, most of this data is unstructured and not suitable for modern bioinformatics methods (fig. 1.1). An example would be one of the most commonly used disease annotation databases, the Online Mendelian Inheritance in Man (OMIM), where genes are annotated with largely unstructured text which is of a historical narrative design, making it difficult to compare or integrate with annotations from other sources. Such unstructured annotation is informative to human user, but it is not computationally friendly, and of limited use in its raw form for model bioinformatics analysis.

Efforts have been made to establish controlled vocabularies for use in gene annotation such as the International Classification of Diseases (ICD) [8], the National Library of Medicine Medical Subject Headings (MeSH) [9] and Thesaurus like the Unified Medical Language System (UMLS) Metathesaurus [10] which were developed to facilitate gene annotation and the building of electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information, as well as in informatics research. The MEDLINE/PubMed database, one of the most heavily used medical databases, is indexed by MeSH controlled vocabularies which greatly improves searches of millions of publication entities.

On the other hand, ontology, sometimes used as terminology, is a formal, explicit representation of a body of knowledge, within a given domain. An ontology provides explicit definition of concepts and their relations as well as cross references to other ontology, features that makes it to be one of the great enabling technologies of modern bioinformatics, especially suitable for annotating genes.

One of the most widely used ontologies in the biomedical domain is the Gene Ontology(GO) [11]. The successful development of the Gene Ontology provided a set of standard, consistent, unambiguous and structured terms to annotate genes or gene products. After more than a decade's effort by the GO consortium and numerous research communities, the GO has became one of the most used and well developed ontologies in the biomedical field. This is partly because GO has a good gene annotation coverage [12]. Originally annotation was performed by human experts/curators who read a research paper and assigned the most relevant GO terms to the genes or gene products studied in the paper. This approach requires extensive domain knowledge from the curators and is time consuming and error prone. A wide range of tools

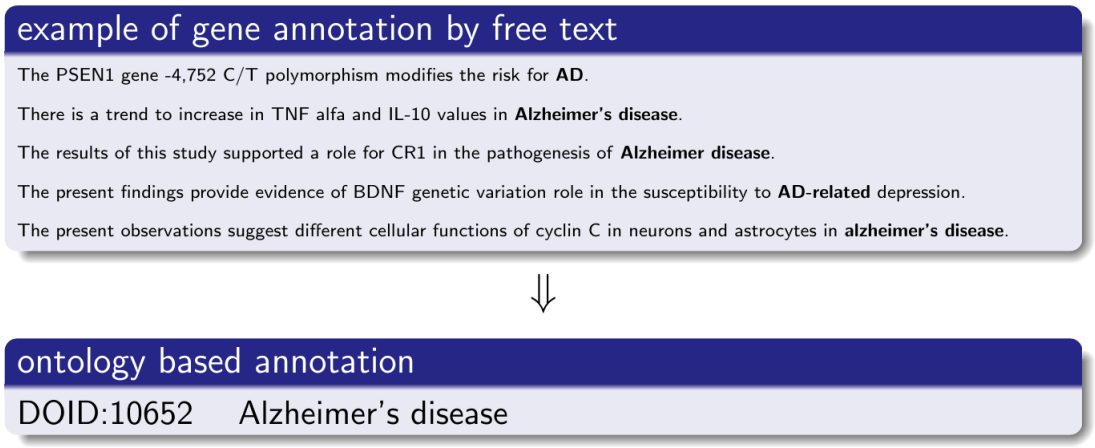


Figure 1.1: An example of unstructured text based annotation and the corresponding ontology based annotation. The five sentences in the top box use five different words (synonyms, abbreviation, etc.) for the disease 'Alzheimer's disease' which make it hard to be annotated accurately, especially when trying to automate the process computationally across large data corpora. Ontologies can provide unique, consistent, formal representations of disease relevant information, allowing reasoning base on their relations, which is neededythe foundation of most of the modern data integration/analysis techniques.

have since been developed to support the annotation process [13]. For example, the Textpresso software tool was one of the first tools developed to support literature curation for GO, and is still used in model organism databases [14]. The Phylogenetic Annotation and Inference Tool [15] is also used by the GO consortium, which assists curators to infer annotations among members of a protein family. Such a 'hybrid' annotation style that combines manual assignment and electronic inference, is helpful to speed up the annotation process, which as a result, produces a better annotation coverage.. What's more, since GO has become a standard in many analysis pipelines, many model organism databases and genome annotation groups use the GO and contribute their annotation sets to the GO resource [11]. The successful development of the Gene Ontology also facilitated a wide range of downstream analytic tools which are reviewed by Huang et.al in [16].

1.1.1 Definition of an ontology

An ontology is a formal, explicit representation of a body of knowledge, within a given domain. Ontologies usually consist of a set of classes or terms with relations between

them. Many definitions of ontology have been proposed in the literature [17–19] based on criteria such as their intended use or degree of formalization. One of the most widely used ontologies in biological sciences, started in around 1998, is the Gene Ontology [20]. By 2007, following the success of the Gene Ontology, interests and demand for ontologies grew and resulted in national and international coordination efforts such as the Open Biomedical Ontologies (OBO) Foundry [21] and the National Center for Biomedical Ontologies (NCBO) [22]. By September 2016, there were more than 500 ontologies stored in NCBO and accessible through the NCBO bioportal web interface. Despite the different definitions, ontologies provide several unique features which are used in almost all of their applications [23]:

1.1.1.1 Classes and relations

Classes and relations between classes are referred to by an identifier. The identifier is consistent across different versions of the ontology, enabling consistent and unambiguous knowledge sharing and data integration. A standard class identifier in an ontology usually consists of a prefix string followed by a colon and a series of digits. In the Gene Ontology, for example, 'GO:0000016' is an identifier for a class while 'part_of' and 'regulates' are relations (fig. 1.2).

1.1.1.2 Controlled domain vocabularies

A set of controlled domain vocabularies, i.e. a list of string labels associated with the ontology's classes and relations used to refer to the kind of things a class or relation represents. They may be provided in multiple languages and multiple labels can be assigned to a single class. A primary label is often used to refer to class while other secondary labels or synonyms are used as complementary labels to capture the usage of a class or a relation in different contexts. The distinction between label and class identifier is that the label may change during the development of the ontology while the class identifier and the intended meaning of the class remains the same. The vocabularies in an ontology aim to cover a domain completely, this usually provides a large set of relevant terms within that domain as well as a set of terms used to describe the ways these terms may interact. For example, GO not only contains classes and relations to represent gene and gene product attributes, but also provides a large set of terms used to refer to these attributes and their relations. An example would be 'maltose biosynthetic process' (the chemical reactions and pathways resulting in the formation of the

disaccharide maltose) has an ‘is_a‘ relation to ‘maltose metabolic process’ and ‘disaccharide biosynthetic process’ while having a set of synonyms including ‘malt sugar biosynthesis’ and ‘maltose anabolism’.

1.1.1.3 Textual definitions and descriptions

These provide additional metadata for classes and relations which provide a precise description of the class. There has been discussion about how to create a ‘good’ textual definition [24]. The majority of ontologies contain two main kinds of additional information, 1) textual definitions and descriptions that provide examples, background information and conditions that make the intended meaning of a class in ontology as precise as possible and 2) additional information that cross link one class to other entries in literature, other databases or other ontologies and vocabularies. For example, ‘A cell cycle process that controls cell cycle progression by monitoring the integrity of specific cell cycle events. A cell cycle checkpoint begins with detection of deficiencies or defects and ends with signal transduction.’ is the textual definition of the GO term ‘GO:0000075 cell cycle checkpoint’ which is cross linked to the Reactome pathway database entities in different species including ‘REACT_100401 Cell Cycle Checkpoints, Gallus gallus’, ‘REACT_1538 Cell Cycle Checkpoints, Homo sapiens’ and ‘REACT_90285 Cell Cycle Checkpoints, Mus musculus’. Another example would be ‘GO:0043076 megasporocyte nucleus’ is a ‘GO:0005634 nucleus’ that is part of a ‘CL:0000320 megasporocyte’ which is a term defined in the Cell Ontology [25].

1.1.1.4 Formal definitions and axioms

‘Machine-readable’ formal definitions and axioms. These are some of the most valuable features of ontologies that enable computational analyses and graph/network based analyses with ontology data. Most commonly, ontologies in biological and biomedical domain are expressed directly in a formal language. The Web Ontology Language (OWL) [26], a formal language based on description logic, has become increasingly popular in representing ontologies. The graph-based OBO flat file format, which was initially used to represent ontologies, has become a sub-language of OWL but tools are available [27] for parsing between the two formats. Currently, in September 2016, 351 ontologies are represented natively in OWL in the NCBO bioportal ontology repository while 106 ontologies are represented natively in OBO.

Graph representation of ontologies can be directly derived from an ontology’s for-

mal definitions and axioms. A typical ontology graph contains nodes and edges where nodes commonly represent classes (terms) and edges represent types of axiom (relations) between classes. An ontology graph is structured hierarchically as a directed acyclic graph (DAG) with a root node. In such as DAG structure, the concept 'level' (depth) was also often used to refer to the location of a node in the ontology structure, for example, the ontology root has a level of 1. It is also appropriate to talk about a parent-child relationship between nodes where parent refers to a connected node closer to the root of the graph, and child to that closer to leaf nodes. The parent (high-level term located near the top of the ontology root) would be a conceptually broader term (class) in the defined domain of the ontology, and the child (low-level term) would represent a more specific term. Nodes can have any number and type of relationship to other nodes as long as there are no directed cycles in the graph. That is to say, a node may have connections to more than one child (more specific) node, or it can also have more than one parent (broader) node, and different relations to its different parents, but it cannot have any connection between its sibling nodes (nodes with the same parent nodes) since it will create directed cycles.

In terms of the type of relationship between nodes, some ontologies including the Human disease ontology and Human phenotype ontology are simpler than others, only using one type of relationship (in most cases, the 'is_a' relationship which can be interpreted as 'is a subtype of'). Other relationships are defined in the OBO Relations Ontology (RO) and used to represent different types of relationship between nodes such as 'part_of' or ' negatively regulates' such as in the Gene Ontology. An important property of the relations between ontology terms is that some of the relations such as 'is_a' and 'part_of' are transitive in nature, which means that new relations can be inferred based on existing ones (dash lines in fig. 1.2a). For example, in the Gene Ontology, the term 'mitochondrion' is an 'intracellular organelle' and 'intracellular organelle' is an 'organelle', therefore mitochondrion is an organelle. This leads to another important feature of ontology hierarchy called the 'True Path Rule'(TPR) which defines each term's meaning by (multiple) inheritance. In other words, the path from a node up to the root node ('True Path') must always be true. Since a node can have more than one parent, there can be more than one true path, thus a True Path Graph of a node is defined to be the sub graph comprising all of the True Paths of the term [28].

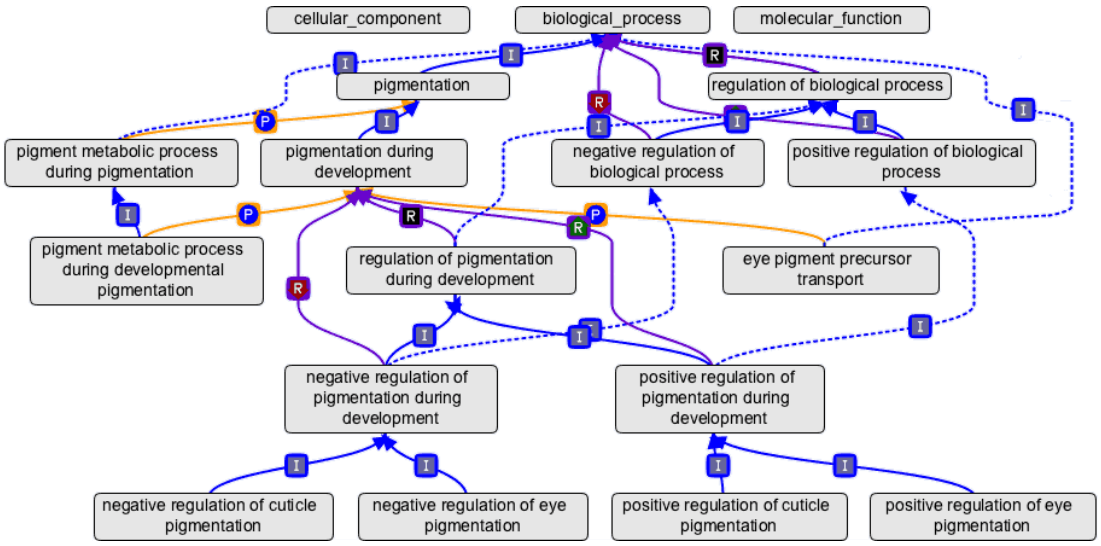
Additionally, annotation from a child term also hold for all of the ancestors in its TPs. This feature is especially useful in the context of ontology based gene annotation: "an annotation for a term in the ontology hierarchy is automatically transferred to its

ancestors based on the TPR, while genes not annotated to a term cannot be annotated for its descendants". Such aggregation of gene annotation inspired the development of new algorithms for a variety of analytic methodologies such as gene function prediction [29–32] and over-presentation analysis [33] which takes advantage of the topology structure of the ontology. The *OntoSuite* framework developed in this project explicitly considers the TPR and will be discussed in later sections of this thesis.

1.1.1.5 Slim/clip ontologies

One common use of the Gene Ontology is functional annotation of results of high throughput experiments, such as transcription profiling arrays [35,36] where GO terms are linked to a particular treatment based on affected genes or gene products. In such cases, the changes in gene expression in response to different treatments can be characterized by a list of associated GO terms rather than a list of genes, which gives additional insights. However, GO includes a very large number of terms, covering most biological knowledge related to gene function from the three main components: cellular component, biological process, and molecular function. For any given system, many terms may be completely unrelated yet are still considered during analysis which usually complicates the analysis and leads to undesired results. In the case of term enrichment analyses, including unrelated GO terms dilutes the proportion of actual hits, increasing the likelihood of falsely reporting enrichment and complicating the interpretation of the resulting list of terms. Consider, for example, the use of the term GO:0007321 sperm displacement, when studying systems such as brain development. Including this term in the enrichment analysis, which in most likelihood has very limited contribution to the interpretation of the result, will increase the number of terms considered in the analysis, in turn reducing the statistical power of the enrichment analysis by requiring more rigorous multiple testing correction.

The most obvious way to overcome this problem is to construct a better defined domain-specific ontology, for example, a brain development ontology and link its terms to all those genes that can be defined by these terms. However, this approach is labour intensive, time-consuming and requires extensive domain knowledge. Moreover, this largely overlaps with the massive efforts of the GO consortium. An alternative approach is to prune existing ontologies such as GO, leaving out the irrelevant terms and only keeping terms that are pertinent to the specific task. This approach has already been used extensively in generating GO 'slims', which is constructed by choosing high-level terms from GO that give a broad overview of the ontology content, such as



(a)

```
[Term]
id: GO:0000016
name: lactase activity
namespace: molecular_function
def: "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose." [EC:3.2.1.108]
synonym: "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
synonym: "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
xref: EC:3.2.1.108
xref: MetaCyc:LACTASE-RXN
xref: Reactome:REACT_100439 "lactose + H2O => D-glucose + D-galactose, Gallus gallus"
xref: Reactome:REACT_104113 "lactose + H2O => D-glucose + D-galactose, Rattus norvegicus"
xref: Reactome:REACT_105850 "lactose + H2O => D-glucose + D-galactose, Bos taurus"
xref: Reactome:REACT_109208 "lactose + H2O => D-glucose + D-galactose, Taeniopygia guttata"
xref: Reactome:REACT_109391 "lactose + H2O => D-glucose + D-galactose, Sus scrofa"
xref: Reactome:REACT_109447 "lactose + H2O => D-glucose + D-galactose, Danio rerio"
xref: Reactome:REACT_112431 "lactose + H2O => D-glucose + D-galactose, Caenorhabditis elegans"
xref: Reactome:REACT_114967 "lactose + H2O => D-glucose + D-galactose, Drosophila melanogaster"
xref: Reactome:REACT_30821 "lactose + H2O => D-glucose + D-galactose, Mus musculus"
xref: Reactome:REACT_78084 "lactose + H2O => D-glucose + D-galactose, Canis familiaris"
xref: Reactome:REACT_78754 "lactose + H2O => D-glucose + D-galactose, Xenopus tropicalis"
xref: Reactome:REACT_9455 "lactose + H2O => D-glucose + D-galactose, Homo sapiens"
xref: RHEA:10079
is_a: GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds
```

(b)

Figure 1.2: (a) Part of the Gene Ontology structure taken from [34]. The domains that GO represents are biological processes, molecular functions and cellular components which are three separate ontologies with no overlapping terms. The nodes in the graph represent ontology terms while the edges represent relation between terms. The edge label indicates the type of relations including ‘I’(is a), ‘P’(part of) and ‘R’(regulates). The dashed lines represent inferred relations based on the ‘True Path Rule’. (b) A GO term definition in the OBO format.

'metabolism' or 'signaling', without the detail of the specific fine grained terms. Currently (September 2016), the GO consortium maintains 10 GO slims including a 'Yeast slim', a 'Metagenomics slim' and a 'Chembl Drug Target slim'. These slims contain a small number of terms that separate gene products into very broad categories within the domain. However, while useful for achieving a bird's eye view, the massive loss of resolution greatly reduces the ability of GO-slims to pinpoint relevant processes, limiting their usage.

Alternatively, instead of using a 'top-down' approach, Geifman et. al. [37] presented an innovative way that prunes the ontology from the 'bottom-up' to produce a domain-specific ontology named 'NIGO' (The Neural/Immune Gene Ontology) that is a subset of the Gene Ontology. "Clipping" selects only the most relevant terms (bottom most term in the ontology DAG) to a specific system and clips irrelevant terms from the ontology. The fundamental difference between a clipped ontology and a slimmed ontology is that, in terms of enrichment analysis, a slimmed ontology will achieve an improvement of statistical scores assigned to the enriched terms which are usually general top-level terms that will reveal a lot about the nature of the biological differences between two sets of samples, but little regarding the specific responses. On the other hand, the clip ontology not only improves the statistical scores but also provides enriched terms from all hierarchical levels which reveals more about the biology underlying the study. In short, better and more comprehensible/interpretable results for functional analysis of microarray data can be achieved, with minimal loss of resolution using a clip ontology. As an example, Geifman et. al. [37] used nine neural and/or immune related microarray data sets together with three non neural/immune data sets(as control) to perform functional analysis with GSEA algorithm [38] on 'NIGO', full GO and a generic GO slim. The results show that functional analysis of neural/immune related microarray experiments with 'NIGO' improved the false discovery rate (FDR) values of relevant GO terms in comparison to the full GO and the generic GO slim with minor loss of relevant terms for the related experiments, but not for neural/immune unrelated experiments. Note that using a clip in analysis such as enrichment analysis is not going to help to generate more significantly enriched terms, on the contrary, some of the significant terms may not pass the threshold when using a clip ontology, others may received reduced significant values due to the reduce of number of irrelevant terms compare to using the full ontology. Thus, the use of a clip can be considered as an extra step to filter out false positives (conservative) from the enrichment result.

Despite the fact that an ontology slim or a clip ontology require less effort than cre-

ating domain specific ontology from scratch, it is still challenging to automate a such process. Even through a 5-step filtration process was proposed to automate most of the work in creating the NIGO in [37], domain knowledge is still extensively needed, thus the process cannot be easily adapted to create clip ontologies in other domains. An alternative method of creating a clip for a domain would be to analyze the enrichment results of a well studied gene list that are relevant to the domain and pick those enriched terms to 'seed' of the clip and walk up/down the ontology DAG from the seeds to manually pick appropriate terms. However, such a process requires human involvement, and is therefore hard to scale and standardize.

1.1.1.6 Essential elements of an ontology

Ontologies provide rich features associated with their controlled vocabularies to represent concepts and their relation to a domain. The minimum requirements of an ontology are a unique identifier and term name, and the definition of terms and their relations. Beside these essential elements of the ontology, some other optional elements that provide extra information are often used, including:

- Secondary IDs or alternate IDs, that refer to a term. These IDs are created when two or more terms are identical in meaning, and are merged into a single term. All terms IDs are preserved so that no information (for example, annotations to the merged IDs) is lost.

- Synonyms. Alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope. This can be very useful when using an ontology as a dictionary for NLP task(see below section).

- Database cross-references, or dbxref, refer to identical or very similar objects in other databases or other ontologies. For instance, the molecular function term "*retinal isomerase activity*" is cross-referenced with the Enzyme Commission entry EC:5.2.1.3; the biological process term sulfate assimilation has the cross-reference MetaCyc:PWY-781.

- Comment, including any extra information about the term and its usage.

- Subset, indicates that the term belongs to a designated subset of terms, e.g. a slim or a clip.

- Obsolete tag, indicates that the term has been deprecated and should not be used.

1.1.1.7 Ontology annotation

Gene annotation refers to the process of assigning relevant information to gene or gene products. Such information can be the gene's functions, relevant diseases/phenotypes or pathways that generated from different research groups, databases etc. The lack of a standard and of guidance for preparing sources and using this information results in a huge variation in the format, granularity and quality of the annotations. For example, disease annotation for 'Alzheimer's disease', in the OMIM database for 'Alzheimer disease' refer to the disorder and contains more than 20 disorder entities with different subtypes ('Alzheimer disease 1, familial' or 'Alzheimer disease 6') and conditions ('Alzheimer disease, type 3','Alzheimer disease, type 3, with spastic paraparesis and apraxia' or 'Alzheimer disease, type 3, with spastic paraparesis and unusual plaques'). In the Ensembl variant database, both 'Alzheimer disease' and 'Alzheimer's disease' are used to refer to the same disease with more than 30 different text entities representing subtypes or conditions including 'Alzheimer disease type 1','ALZHEIMER DISEASE FAMILIAL 3', 'ALZHEIMER DISEASE FAMILIAL 3 WITH UNUSUAL PLAQUES' and 'Alzheimer's disease (late onset)'. The two databases use different terminologies for the same disease, and even the same subtypes are represented differently ('Alzheimer disease type 1' vs 'Alzheimer disease 1, familial'). Such inconsistency and unstructured text based annotation is far from comprehensive, not computation friendly and hard to be integrated or used in any modern analysis technique.

1.1.2 The challenges of ontology based annotation

The Gene Ontology (GO) (fig. 1.3) based annotation is currently the only annotation data source used the most, due to the fact that the GO has been developed for over a decade to get a good annotation coverage. GO is designed to represent knowledge for biological process, molecular functions and cellular components. However, annotation with single ontologies is often not sufficient to explain experimental results. There are many other established and emerging ontologies that would be beneficial for the biological interpretation of data in a different range of areas but they are rarely used. For example, there were more than 500 unique ontologies (September 2016) in the NCBO repository [22], including the Human Disease Ontology (DOID) [39], Human Phenotype Ontology (HP) [40] and Alzheimer's disease ontology (ADO) [41], each of which

focuses on a specific domain. Together, these ontologies can be used to formalize large, unstructured and often diverse data types such as those from clinical records. Multiple ontologies can be used together to capture information and knowledge from different perspectives including drug, phenotypes, disease, environmental factors, behavior and diagnostic procedures, which will potentially lead to novel therapies, methods for patient classification and clinical decision making, which is one of the key objectives of TBI. However the lack of annotation coverage in these ontologies greatly limits their usefulness and places a huge barrier to their usage with existing data analysis techniques [42]. As an example, the gene HMOX2 was shown to be involved in the process of pigment biosynthesis by McCoubrey et al. in 1992 [43] but it is still not annotated as such today. More commonly, because of the time lag necessary for the manual curation process, some of the recent discoveries are not yet available.

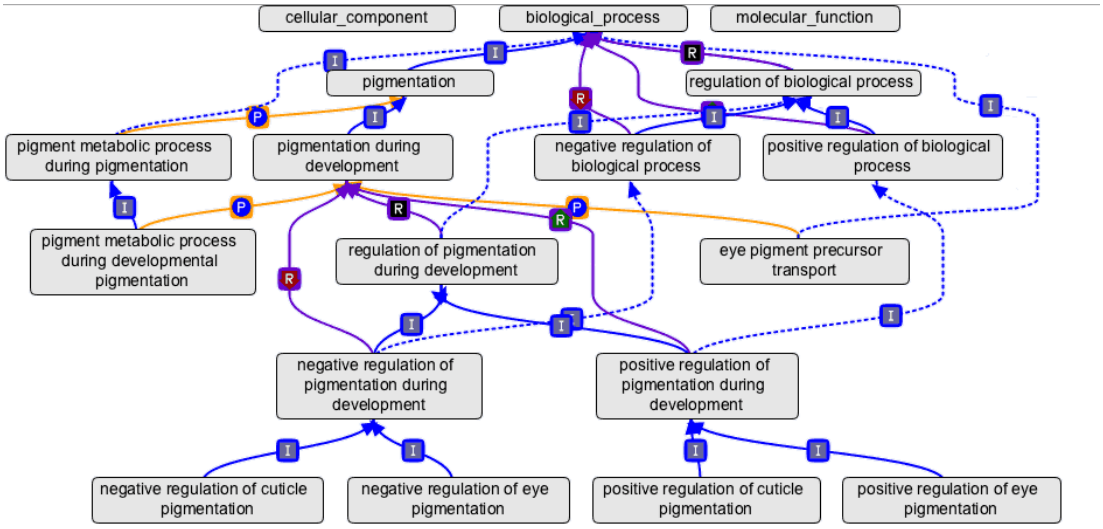


Figure 1.3: Part of the Gene Ontology structure. Nodes represent Gene Ontology terms while edges between nodes indicate their relationship. The most common relation is the 'is_a' relation, indicated by a capital 'I' on the edge, which connects a GO term with its sub concepts(children terms). Other relation including 'R' for regulates and P for 'part_of' are also defined in GO.

Since manual curation is not sufficient [44], it is impractical to manually annotate all ontologies. In recent years, Natural language processing (NLP) has been proven to be a reliable and accurate method to identify and extract relevant data from biomedical text corpora, showing promise for the automation of the ontology annotation process [45]. Attempts have been made to use text mining to learn more about human diseases. Rappaport et. al. [46] proposed a disease annotation system called 'MalaC-

ard', which integrated 44 disease sources and used an in house automatic disease name unification algorithm to generate disease entities. Text mining algorithms have also been to improve chemical-gene-disease curation in The Comparative Toxicogenomics Database (CTD) [47] and the 'Genetic Association Database'(GAD) [48] which aims to collect, standardize and archive genetic association study data with data mining tools. 'DisGeNET' provides gene disease association data by combining manually curated data together with an NLP-based approach [49]. These works all use text mining approaches to discover gene disease associations from biomedical text, however, instead of using ontologies to represent diseases, they use different terminologies for representation/classification. This lack of consistency on the use of formal data structures makes it hard to determine the number and types of diseases each database contains. It also makes it impossible to compare/integrate gene disease annotations across these databases. For example, in the GAD database, a search for 'Crohn's disease' returns 25 results but searching for 'regional enteritis' returns no results.

Osborne et. al. [3] proposed an NLP based approach with the MetaMap [50] and mined the NCBI GeneRIF database for human gene disease associations. Instead of using unstructured terminologies, Human Disease Ontology (HDO) [39] was used to represent human diseases. The resulting HDO based gene annotation was assessed against the Homayouni gene collection [51] and suggested a 91% recall rate and 97% precision rate. In a similar approach, LePendu et. al. [4] applied a different NLP approach with the NCBO annotator [52] and mined HDO terms from titles and abstracts of PubMed entries. The resulting HDO based gene disease annotation can be easily reasoned, aggregated, filtered, and cross-referenced. Such ontology based annotation facilitates a range of downstream analysis, for example, ontology term enrichment analysis and network analysis, which are likely to bring new insights into human diseases.

1.2 Existing Tools for Functional Annotation Analysis

The second part of the TBI challenge is the availability of the appropriate bioinformatics tools for the analysis and exploitation of gene annotation data. Functional analysis of gene lists, derived in most cases from high-throughput genomic, proteomic and bioinformatics scanning approaches, is still a challenging and daunting task. Over the last decade, gene-annotation enrichment analysis has become standard practice in the analysis of such lists, making it possible to systematically assess enriched and pertinent

biological features and processes. The assumption underlying enrichment analysis is that biological processes are the result of the cohorts of genes rather than a single individual gene. Under this assumption, for a particular biological process in a given study, the co-functioning genes should have a higher (enriched) potential to be selected as a relevant group by the high-throughput screening technologies. Instead of focusing on an individual gene, such analysis consider a group of relevant genes at the same time, increasing the likelihood for researchers to identify the most pertinent biological processes under study.

Many Bioinformatics enrichment tools have been developed over the last decade. These tools play an important and successful role in gene functional analysis and are often applied to gene lists for various high-throughput biological studies [49, 53–59]. In a survey, Huang et al. [16] identified and reviewed at least 68 different enrichment methods. The 68 methods were classified into three categories: singular enrichment analysis (SEA); gene set enrichment analysis (GSEA); and modular enrichment analysis (MEA).

Singular Enrichment Analysis (SEA)

SEA is the most traditional way to identify predefined annotated gene sets (ontology terms such as GO terms in ontology based annotation), that appear significant more frequently than random in a candidate gene list. Such enrichment analysis often takes a list of preselected 'interesting' genes (e.g. differentially expressed genes with a p-value $\leq$0.05 and fold-change $\geq$1.5) and then iteratively tests the enrichment of each ontology term one-by-one in a linear mode. The resulting enriched ontology terms are then ranked by significance and presented as a flat list. A toy example of the SEA approach is shown in fig. 1.4.

The SEA approach is a simple but efficient way to gain a first biological insight into the important functions associated with a set of genes. A family of enrichment tools was developed, implementing the SEA approach, such as GoMiner [55], OntoExpress [54], DAVID [60], GIEASE [61] and GFinder [62]. They have been widely used in enrichment analysis with significant success by many researchers [49, 56–59]. However, since the pre-selection genes for the input list has a great impact on the enrichment result. For example, the use of differentially expressed genes with a p-value $\leq$0.05 vs a p-value $\leq$0.01 from the same experiment may result in totally different enrichment result. Such arbitrary 'cut-off' made to select the genes makes SEA subjective to a certain degree. The SEA is also criticized for its assumption of the indepen-

dence of genes within the predefined gene sets (genes annotated to an ontology term). In fact, an extensive correlation between genes is a well-documented phenomenon. For example, Gatti et. al. [63] investigated the correlation between genes within gene sets (genes annotated with the sane GO term and genes involved in the same KEGG pathway) by analyzing 8,656 arrays from Gene Expression Omnibus [23] data. Their results gave strong evidence of consistent correlations between genes in GO term and KEGG pathway, suggesting that the gene independence assumption is inappropriate and the correlations between gene within gene set are non-trivial which, if ignored, may lead to overly optimistic results (smaller p-value with more false positives) in SEA analysis [16, 64–67]. In addition, the flat list of enriched terms generated by the SEA can something be very large and overwhelmed by groups of 'similar' terms, for instance, Gene Ontology terms like cell growth, unidimensional cell growth, multi-dimensional cell growth, cell tip growth, pollen tube growth, etc., which dilute the interrelationships of relevant biology concept in the result.

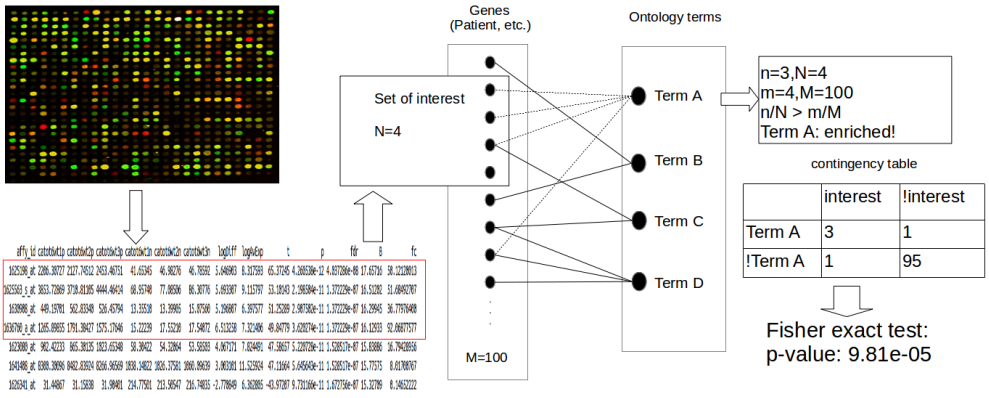


Figure 1.4: A toy example illustrating the SEA work flow: First, a candidate gene list of interest is generated. This, for example, could be a list of genes from a microarray experiment where gene expression values are measured in different conditions. A p-value is given to each gene based on a null hypothesis and those genes that passed a specific cut-off will become the candidate genes. Then, for each ontology term, the frequency of the term in the candidate gene list is compared to the frequency of the term in a reference background (usually includes all the ontology terms and their annotated genes). Finally, A two-by-two contingency table is created and used to assess the significance of such an observation, accounting for the size of the candidate gene list and the background, for example, using a Fisher exact test. The background set in the example consists of M=100 genes which is all the gene from the microarray with at least one ontology annotation. The set of interest has N = 4 genes which are the genes that pass the user defined threshold (p-value $\leq$0.05 and fold-change $\geq$1.5). Term A has three links to 3 genes within the set of interest (n = 3) and 1 genes to the background set which is not in the set of interesting (m = 4). A contingency table was created to calculate the statistical significance with Fisher's Exact Test, which in this particular example, $p - value = 9.81e - 05$, indicating a significant enrichment of Term A

Gene Set Enrichment Analysis (GSEA)

To overcome the arbitrary 'cut-off' limitations of the SEA approach, gene set enrichment analysis (GSEA) was introduced in [68]. The GSEA is a 'cut-off' free approach that takes all the genes from a high-throughput experiment, for example at the same time. The primary advantages of GSEA are that 1) it objectively considers the entire list of genes so that those genes that would have been removed by the 'cut-off' (with a relatively small fold-change or ranked at the bottom of the gene list) are included and contribute to the enrichment analysis in certain degrees; 2) it used a permutation based resampling approach to estimate the enrichment, maintaining the gene-gene dependency that reflects real biology and have been proven reducing the false positives from the result [63, 69].

In a typical GSEA scenario, genome wide gene expression profiles are generated from samples belonging to one of two classes, for example, tumors that are sensitive vs. resistant to a drug. These genes are ranked by their correlation to the classes C with a ranking metric, for example, signal to noise ratio. The task is to find out whether a predefined gene set S (genes in the same pathway or sharing the same GO annotation) is overrepresented towards the top or bottom of the ranked list L. An enrichment score (ES) is calculated by walking down the ranked list L, increasing a running-sum statistic when encountering a member gene of set S (P_{hit}) and decreasing it when encountering a non-member gene of set S (P_{miss}) (Equation eq. (1.1) on the following page). The ES is defined to be the maximum deviation of the running-sum from zero across all genes ($P_{hit} - P_{miss}$), which corresponds to a weighted Kolmogorov-Smirnov statistic. The magnitude of the increment during the running-sum process is controlled by an exponent p. When $p = 0$, increment steps are equal to those in a standard Kolmogorov-Smirnov statistic. When $p = 1$, the increment of hitting a gene in S is weighted by their correlation to class C normalized by the sum of the correlations over all of the genes in S. For each gene S, a permutation test is performed to shuffle the gene expression values a predefined number of times N in order to obtain an ES_{null} distribution. An enrichment p-value for S is then computed as the fraction of shuffles which produces an ES at least as great as the observed.

The Leading-Edge Subset of genes in gene set S, as defined in [38], are those genes in the gene set S that appear in the ranked list L at, or before, the point where the running sum reaches its maximum deviation from zero (fig. 1.5). These genes can be interpreted as the core members of a gene set that account for the enrichment signal and these often reveling important biological insights as shown in [38].

$$P_{hit}(S,i,j) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \text{ where } N_R = \sum_{g_j \in S} |r_j|^p, P_{miss}(S,i,j) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)} \quad (1.1)$$

where $P_{hit}(S,j)$ is the normalized step length for the j th gene in the ranked List L which also in the gene set S, $P_{miss}(S,i,j)$ is the normalized step length for the j th gene in the ranked List L but not in the gene set S. i is the position in order list L, r_j is the correlation of gene g_j with class C, N is the total number of genes in L and N_H is the number of overlapping gene in S and L.

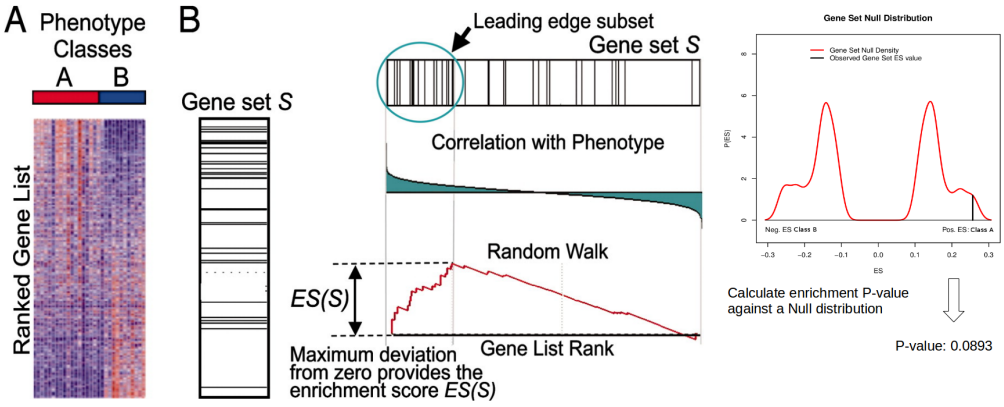


Figure 1.5: GSEA work flow. Genes are ranked by their correlation to the classes (e.g. disease vs control) with a ranking metric, for example, signal to noise ratio. An enrichment score (*ES*) is calculated by walking down the ranked list *L*, increasing a running-sum statistic when encounter a member gene of set *S* and decreasing it when encounter a non-member gene of set *S*. The *ES* is defined to be the maximum deviation of the running-sum from zero across all genes, which corresponds to a weighted Kolmogorov-Smirnov statistic [38].

In the original implementation of GSEA [68], the exponent p was set to 0 which yielded high *ES* value for gene sets clustered near the middle of the rank list. These sets were found not to represent biologically relevant correlation with the classes. This issue has been solved by an improved version of GSEA introduced by Subramanian et al [38] which weights the steps based on each gene's correlations with the sample classes (i.e. condition vs control or A and B in fig. 1.5), which correspond to seting $p = 1$. In this project, a conceptually similar GSEA algorithm is proposed and implemented. The exponent p has been set according to a confidence score based on *HDGDB* which increases the magnitude of the effect of those confident annotations

and decreases the magnitude of the effect for less confident ones. This will be discussed in the detail in section 3.2.7.

Since the publication of the original GSEA method, other GSEA-like gene set analysis strategies have been developed and implemented such as SAM-GS [70], LRpath [71] and GAGE [72]. Tools like ErmineJ [73], GO-Mapper [74], ADGO [75] use slightly different statistical tests but keep the 'cut-off' free feature of GSEA. However, a common feature of GSEA is that the enrichment p-value is mainly driven by genes towards the extreme of the ranked list (top or bottom, usually those with the highest fold change). These genes being highly weighted, contribute the most to the ES score. This is not always true in real biology since some big changes in a gene's expression may be caused just by some small but important signal regulation events. Depending on the research scenario, subtle changes in gene expression can be just as important. In addition, GSEA is not suitable for experiments that only have small number of samples. This is because the enrichment p-value is calculated base on a permutation test that shuffles the sample labels to obtain a 'null' distribution. A small sample size heavily affects the quality of the 'null' distribution, thus yielding an unreliable enrichment p-value. For example, if the experiment generates 6 samples belonging to 2 classes, then the estimated p-value is always greater than 0.05. When only 4 samples are available, the estimated p-value is always greater than 0.16, which will not generate any significant results with a commonly used 0.05 threshold. In practice, a minimum of 10 samples is usually required for effective GSEA which is not always possible in real biology experiments due to the fact that these experiments are often time consuming and expensive. More evidence on how enrichment results vary depending upon the choice of 'null' distribution can be found in [76]. A workaround to use GSEA on small sample size data is, shuffle the gene labels in the permutation procedure rather than the sample labels to obtain the 'null' distribution. The rationale of this approach is that a significant gene set should be distinguishable from an equally sized set composed of randomly chosen genes. However, gene label shuffling is not strictly appropriate because it breaks the gene-gene correlations, thus resulting overestimation of significance level. This method can however, be useful for hypothesis generation. Sample label shuffling is generally favoured because it preserves the relationship between genes and addresses the question of identifying S whose expression changes correlate with sample class changes. The differing two shuffling approaches generate different 'null' distribution which in turn, often leads to different conclusions [77]. Last, but not least, the KolmogorovSmirnov statistic requires a relatively large number of data points to

properly calculate the p-value. Thus the GSEA method works better on a predefined gene set *S* when *S* contains a large number of genes. It's performance and reliability drop as *S* becomes smaller in size.

Modular Enrichment Analysis (MEA)

MEA is an extension of SEA/GSEA where network discovery algorithms are considered base on term-to-term relationships among annotations. Two main kinds of MEA currently exist. The First kind of MEA uses clustering algorithms to either pre-process the referenced background or post-process the enrichment result from SEA/GSEA. For example, tools like ADGO [78] and ProfCom [79] implement algorithms to pre-process ontology based annotation, using boolean set operations including intersection, union and subtraction, to generate composite (joint) annotation terms which are used subsequently as the reference background in the enrichment analysis. These algorithms were motivated by the insufficiency of single ontology in explaining the changes of specific expression patterns. For example, not all of the genes categorized by a biological process may alter their expressions as a whole under an experimental condition, but only those with a particular localization or those involved in a particular pathway might alter their expressions. Thus, if genes of term A from Gene Ontology Biological Process (GOBP) overlaps with term B from Gene Ontology Cellular Component (GOCC) to a certain degree (can be adjusted accordingly), a set union is performed and a composite term is created with genes involved in a particular biological process (A) which takes place in a specific location (B). It is suggested that the use of such joint annotations can improve discovery sensitivity and specificity in enrichment analysis [78]. Other tools like COFECO [80]and GENECODIS [81] implement an association rule-mining algorithm to extract co-occurring annotations and apply clustering algorithms on the enrichment results to group 'similar' terms into the most relevant meta groups. These approaches take advantage of the relationship between ontology terms and create joint terms that may contain more meaningful biological information than individual ones. The results are highly redundant and have interrelationship regarding different aspects of the underlying biology being studied.

The Second kind of MEA makes use of the ontology hierarchy structure to improve the design of the enrichment algorithm. Ontology terms are typically structured as a directed acyclic graph (DAG), with nodes being the terms and edges being the relations ('is_a', 'part_of', etc.). Nodes towards the top (high) of the hierarchy represent general terms (e.g. a cell) where nodes become more and more specific as you travel down

the DAG towards the leaves (e.g. synaptic vesicle). TopGO [33] developed a gene elimination method which utilizes the topology of the ontology structure to find the most specific term during enrichment analysis. The algorithm starts testing enrichment from the term located at the bottom of the hierarchy, removing gene annotations from all of the parent terms if a term is found to be significantly enriched, otherwise, the gene annotation is rolled up to the parent terms. This effectively stops gene annotations contributing repeatably to higher level terms (more general) thus favouring reporting of the most specific terms. A similar approach has also been developed in the pathway analysis field where the pathway topology structure is used to improve analysis [82].

The key advantage of the MEA approach is that it implements the basic concepts of SEA/GSEA while incorporating term-to-term relationships shifting enrichment analysis from term-centric to biological module-centric. By taking into account the redundant and graph-structure of ontology annotations, respects the fact that biology often works in a co-ordinated manner, a bigger and more meaningful biological picture may be generated from these result. However, MEA implements the core of the SEA/GSEA and so also inherits their limitations. For example, the quality of the pre-selected gene list impacts on the results, just as it does in SEA analysis. In addition, orphan terms or genes are likely to be omitted which could have brought important insights to the analysis.

Despite the distinct features of the different enrichment tools, the general procedure for enrichment analysis is similar and can be summarized into three parts: 1) preparation of the backend annotation database; 2) calculation of the enrichment (algorithm and statistics) against a reference set; and 3) post-process and presentation of the result. It is a common misunderstanding to only consider the statistical method alone in enrichment analysis. In fact, each of the three parts has great influence on the final enrichment results. Based on these three parts, some common limitations are revealed.

1. Backend Annotation Database

In general, enrichment analysis is designed to detect significantly over-represented annotations shared by a set of interesting genes when compared to a reference background gene set. The quality of the annotation is one of the most important components in the analysis process and has a great effect on the results. In general, two types of annotation exist in the biological domain, unstructured and structured. The former usually refers to annotation that uses plain text while the latter usually refers to ontology based annotation. Huge variation and duplication exist in unstructured annotations due

to the fact that they are created by different experts using inconsistent, non-standard terminologies and are not suitable for quantitative analysis such as enrichment analysis. On the other hand, ontology based annotations provide a formal and consistent vocabulary for representation of domain knowledge. The structure of the ontology naturally reflects the interrelationship between the ontology terms, respecting the true underlying biology, which is very important especially for MEA algorithms that take into account such structure when calculating the enrichment. In addition, ontology based annotation is amenable to computer manipulation which makes them ideal for use in enrichment analysis.

The Gene Ontology (GO) [20] based annotation is currently the only annotation data source used in most, if not all of the enrichment analysis tools because GO provides a relatively good gene coverage to be used as a reference background. However, there are many other established and emerging ontologies that would be beneficial for biological interpretation in different aspects but are rarely used in enrichment analysis due to the lack of annotations [16]. In addition to incomplete annotations, some of the existing annotations are inaccurate. For example, out of 481685 total GO annotations available for *Homo sapiens*, 155499 (32%) are inferred exclusively from electronic annotations (with Evidence Codes *IEA*, no human expert has been used to check the annotation's accuracy). Even though the vast majority of them are reasonably accurate [83], some are incorrect [84, 85].

In terms of disease, a recent survey by Rappaport et al. [46] identified more than 60 disease-related databases, each of which focuses on different aspects of disease annotation and/or contains a specialized list. Moreover, different biological aspects are being maintained and annotated by different independent resources which have different focus. For example, OMIM [86] contains gene-disease associations while KEGG [87] mainly focus on pathways; BIND [88] stores protein-protein interactions while protein domain information can be found in Pfam [89]. A comprehensive backend annotation data set should integrate diverse and heterogeneous data sources in a coherent way to provide a more reliable reference rather than using single data sources such as GO alone.

Despite the advantages of using ontology based annotations, the usefulness of enrichment analyses is impacted by the annotation bias present in ontology based annotation databases. Some biological processes or diseases are studied in more detail than others, thus more data are available for the corresponding ontology terms. If more data about a specific ontology term is available, more of the genes associated with it will be

known and hence, the term is more likely to appear as significant than others during enrichment analysis.

I discussed in chapter 2 on page 33 how we improve the backend annotation database by implementing a data integration framework and integrated three publicly available data sources including GeneRIF, OMIM and Ensembl variation using the Human Disease Ontology. The resulting annotation database, named *HDGDB*, is potentially more useful than any of the three single databases and is used throughout subsequent work in this thesis.

2. Creating an appropriate gene reference background

As noted in the previous example, if 20% of the genes under study are found to be associated with the *synapse complexity* compared to 8% of genes in the human genome as a while, a statistical test could be performed against a null-hypothesis passed at a user-selected α value. In this example, the genes under study are compared to a reference gene set, the human genome. The choice of the reference gene set has a great impact on the calculation of p-value, even when using the same statistical test.

Two main approaches exist for selecting reference gene sets. The first approach selects reference genes background based on the available genes, that is all genes from a genome [55,90,91] or, for example in a microarray experiment, to select those genes that exist on a microarray [53, 54], since a gene that is not on the array can never be found to be differentially expressed , thus including those genes will only increase the size of the background, which often result in an overly optimistic $p-value$. A second approach of selecting reference genes is, on top of the first approach, based on the annotation availability. For example, when using GO, [90] uses all the genes that have a GO annotation as the reference gene set. Recently this method has been improved by selecting genes only with annotations that are relevant to the study. By utilizing the ontology structure, a sub-set of the ontology terms are first selected to form a pruned version of the ontology that is specific to a narrowed-down domain, a 'clip ontology' (see section 1.1.1.5 on page 7) in [37]. Then the genes with annotations from this sub-ontology are selected as the reference gene set. This method reduces the number of genes in the background, thus tend to be conservative but has good performance in eliminating false positive as reported in [37].

There is no 'gold' standard method for selecting a gene reference background. Sometimes biology experts have domain knowledge that can be used for guiding the selection of the reference genes. When this information is not available, I believe that

	Null hypothesis is True H_0	Alternative hypothesis is True H_1	Total
Declared significant	V	S	R
Declared non-significant	U	T	m-R
Total	m_0	m-m_0	m

Table 1.1: Various errors committed when testing multiple null hypotheses. V is the number of false positives (Type I error) (also called 'false discoveries') while 'T' is the number of false negatives (Type II error).

using genes with relevant annotations is the best approach.

3. Multiple Hypothesis Testing

In a typical enrichment analysis, a list of interesting genes are tested against a number of predefined gene sets (ontology terms) simultaneously against a null-hypothesis passed at a user-selected α level (commonly set to 5% or 1%). Those terms that have a p-value less than the pre-selected α level are considered significantly enriched. However, according to statistical principles, the more gene sets that are tested at the same time, the greater the chance of an increase in the Family Wise Error Rate (FWER, or type I error, 'V' in table 1.1, the probability of making one or more false discoveries among a family of hypothesis tests) [92, 93], thus, any of those significantly enriched terms can actually appear with a non-zero probability just by chance. This issue, refers to the multiple hypothesis testing problem, which is well recognized in the field [55, 90, 94–97]. The un-adjusted p-value can be misleading and correction is need. table 1.1 shows possible errors committed when testing m null hypotheses. It defines some random variables that are related to the m hypothesis tests. m_0 is the number of true null hypotheses. V is the number of false positives (Type I error, also called 'false discoveries'), R is the number of rejected null hypotheses ('discoveries'). FWER can be formally defined as $FWER = Pr(V \geq 1)$.

Methods are available to control the FWER in multiple hypothesis testing. Bonferroni correction is one of the simplest methods used to counteract the problem of multiple comparisons. It controls the FWER by dividing the un-adjusted p-value by the number of tests performed to ensure that the probability of making even one type I error in the family stays less than a certain level (such as $\alpha \leq 0.05$). This method is widely used, but is known to be overly conservative in the sense that while it reduces the number of false positives, it also reduces the number of true discoveries. Thus it is not suitable for analysis where many gene sets are involved (e.g. more than 50) which

unfortunately is the case for most enrichment analyses. Improved methods including Holm's step-down procedure [98] and Hochberg's step-up procedure [99] were introduced as less conservative adjustments of the p-value. For example, in Holm's step-down procedures, p-value are ordered from lowest to highest. For a given level α, the procedures walk down the list of p-value, accepting the test until the smallest k_{th} such that $P_k > \frac{\alpha}{m+1-k}$, where m is the number of tests performed and k is the index in the ordered p-value list.

However, these statistical methods assume independence amount the individual test, which is known to be false for most of the ontology based enrichment analysis. For example, the hierarchy of the Gene Ontology indicates that many terms are closely related, sometimes as parent-child, sometimes as siblings sharing the same parents. In addition, methods that control the FWER are often criticized as having low power (the ability of a test to detect an effect). These methods are suitable for studies where any false positives can lead to a large waste of time to experimentally tested, but are way too conservative for others such as gene expression data generated by high-throughput technologies where small sample numbers are tested and large numbers of variables are measured. Guarding against one or more false positives is typically going to be too strict and will lead to many missed findings of interesting hypothesis.

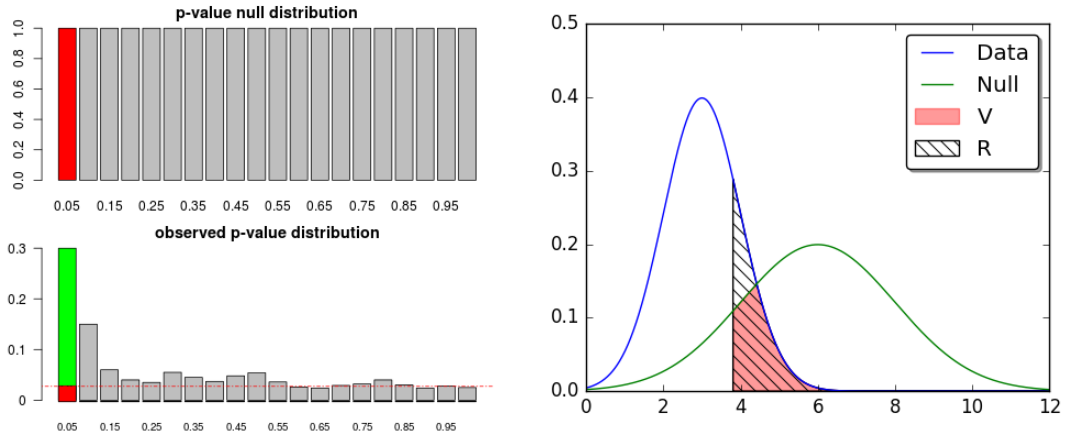
A better question to ask, is how many errors are expected, the so called false discovery rate (FDR,the proportion of type I errors among all rejected hypotheses) [93]. FDR is more flexible and has better power than FWER, which is particularly good for exploratory analysis where it is better to have mostly true findings, rather than guarding against one or more false positive results. FDR can be formally defined as $FDR = E[\frac{V}{V+S}] = E[\frac{V}{R}]$ (see table 1.1 for notions). Controlling the FDR is a process that controls the distribution of the test statistics,f, for example, the distribution of p-values. f is generally considered as the mixture density of the two populations [100]:

$$f(x) = \pi_0 f_0(x) + (1-\pi_0) f_1(x), \pi_0 = \frac{m_0}{m} \tag{1.2}$$

Where f_0 and f_1 are the distribution of the test statistic under the null-hypothesis (H_0) and the alternative hypothesis (H_1) respectively, m_0 is the number of true H_0, m is the total number of hypotheses and π_0 is the proportion of true H_0 among m.

Methods such as those proposed in the original FDR paper by Benjamini et al [93], compute the FDR directly from the p-value, without estimating π_0 (equal to $\pi_0 = 1$, assuming complete H_0), providing the strongest control on FDR but with the lowest power. Alternative methods like those in [101–115] were proposed to estimate

π_0 and control FDR based on an estimated π_0 (Figure 1.6a). However, these estimations are based on the assumption that f_0 is uniformly distributed. This is not true in datasets with large scale strong correlations where the observed f_0 severely deviate from uniform, causing π_0 estimation methods to become very unstable which in turn makes FDR estimation unreliable [107]. The effect of correlation on simultaneous significance tests was previously studied theoretically in [116–118]. Improved techniques have been developed such as in [67] where re-sampling strategies were used in strongly correlated simulated dataset and the estimations of π_0 is reported greatly improved. After the original FDR paper Benjamini and Hochberg (1995) [93], there were increasingly interests in developing methodologies for controlling the FDR under different model assumptions, instead of the independent assumptions. In a later paper, Benjamini and Yekutieli (2001) (hereafter called BY) [119] relaxed the independence assumption to certain dependence structures, that is, when the underlying statistics are positive regression dependent on a subset of the true null hypotheses. A conservative step-up procedure controlling the FDR was developed to control the FDR in the strong sense without relying on the independence assumption. The BY procedure was used as the default method for multiple hypothesis correction for all the result presented in the thesis.



(a) (b)

Figure 1.6: Two methods for FDR estimation. a) FDR is estimated by assuming a uniform distribution of the test statistic, for example, p-value. The top histogram shows the uniform distribution which is likely to be seen in an experiment where there is no significant changes. Note that even if there are no significant changes in the experiment, it is still expected, by chance, to get p-values$<$0.05 (in red). In the bottom histogram, significant changes are observed (in green), some might be false positives. Based on the uniform distribution assumption, a line can be drawn where the p-value distribution flattens out which helps to estimate how many significant values are actually false positives (the red portion of the green bar). b) FDR is estimated against an empirical null distribution. The empirical null distribution is usually derived by re-sampling or performing a permutation of the dataset. For a given threshold (vertical line), the shaded area represents all data that is considered significant (All discoveries, R in table 1.1) while the overlapping area (in red) represents false discoveries (type I errors, V in table 1.1). The ratio of these two areas is then used to estimate the FDR. (the figure shows the idealized distribution, the actual distribution would usually be depicted with a histogram or bar graph)

Other methods are available to compute the FDR. For example, instead of using the distribution of the test statistics, calculating it directly from the dataset using re-sampling/permutation based algorithms such as SAM [120], dChip [121] and GSEA [38]. In theses methods, the FDR is computed against an empirical approximation of the null distribution (f_0) which naturally considered the correlation between the tests (fig. 1.6b). However, these methods usually require a large amount of sample data to compute a reliable null distribution which is a limitation especially in the context of biological experiments. In addition, like methods by Benjamini et al [93, 119], these permutation based algorithms do not estimate π_0, thus the FDR control is considered to be overly conservative especially in datasets with a large number of true alternative hypothesis (H_1).

Multiple hypothesis correction has been studied for many years and it is, not just in the context of enrichment analysis, but in general still a very active field. It is very important to be aware of such problems when doing multiple hypothesis tests but there is not much evidence of how much of an improvement in discovery sensitivity and specificity can be achieved by applying these methods in real-life practice in enrichment analysis. By comparing various common correction methods that are with real-life datasets in Gene Ontology enrichment analysis, [122] concluded that common multiple testing correction methods that are overly conservative approaches for enrichment analysis involving thousands of annotation terms, may negatively affect specificity rather than improve it. As pointed out by [16], enrichment analysis results are influenced not only by the statistical methods used, but also by the algorithms and data sources used. This can not simply be fixed by multiple testing correction, suggesting that efforts to improve sensitivity and specificity should first be fundamentally addressed before refining statistical approaches.

4. ID Mapping

Annotation databases typically provide annotations for genes using some kind of gene identifiers such as the EntrezGene ID [123] or HGNC gene symbol. High throughput technologies like microarrays use their own probe names to identify specific nucleotide sequences, which map to specific genes. Thus, in order to perform enrichment analysis for a list of differentially expressed genes, an essential first step is to effectively translate the list of probe IDs into a list of corresponding gene IDs that matche the annotation source. For proteins id such as those from UniProt [124], genes need to be further mapped to proteins. The success of such ID-to-ID and ID-to-annotation

mappings have a large impact on subsequent enrichment analyses.

Currently, identifiers for genes and proteins are spread out across a number of databases and other resources and maintained by various independent bioinformatics organizations/groups/companies that often have very different interests and research foci. As a result, these resources often use their own identifiers. Cross referencing these identifiers is still challenging. For example, Entrez Gene does not cover PIR ID while UniProt does not reference RefSeq ID. Even between the most commonly used gene identifiers Entrez Gene and Ensembl Gene, there are not always perfect matches. Such ID mapping issue becomes crucial when mapping from one type of identifier to another is not one-to-one. For example, some Entrez Gene IDs are mapped to multiple Ensembl Gene IDs and vice versa which may cause confusion and lead to incorrect interpretation in the following analysis. What's more, low-resolution mapping results in loss of important information. For example, *XRN2a* , a variant of *XRN2*, is mainly expressed in human tissues, whereas another variant of the same gene, *XRN2b* is found expressed in blood leukocytes [125]. Only *XRN2* exists in the Entrez Gene database with an Entrez gene id of 22803. Tissue-specific information, and possibly condition-specific information, provided by the two variants, will be lost during the ID mapping. Another out-standing issue is to map identifier across model species. This is usually done by using orthologues genes but may suffer from low coverage or high noise (many-to-many mapping) issue discuss previously in section 2.3.5.

The increase of the id mapping can also result in an increase of the annotation content, as reported in [126] that 10-20% more GO terms were able to be assigned to the corresponding genes in DAVID after a process called the DAVID Gene Concept, a single-linkage algorithm to agglomerate redundant gene IDs into the DAVID gene clusters in order to improve cross-referencing capability across several independent database sources, particularly between NCBI and UniProt systems. Even though efforts have been made to improved the ID-to-ID and ID-to-annotation mapping and tools have been developed such as Onto-Translate [54], MatchMiner [127], IDConverter [128] and DAVID ID Converter [129], the ID mapping problem is yet to be fully solved and is still a burden left entirely on the shoulders of the researchers. More effort is still needed from the major bioinformatics organizations to improve the quantity and quality of their cross mapping data.

To sum up, despite the usefulness of these tools and the different statistic algorithms used for finding the enriched ontology terms, most of them only work for a specific type of ontology, GO in the majority cases, even though the general under-

lying enrichment analysis process is similar. The difference in statistical algorithms, parameters setting make it not intuitive to compare the enrichment results generated by these tools under the same environment. A standard framework for general ontology enrichment analysis is needed, which is, alongside the lack of ontology annotation, another limitation for the use of newly emerging ontologies.

Therefore, my main motivation of the work presented in this thesis can be summarized as follows: *Ontologies are semantic frameworks upon which biological data can be structured and have grown to be one of the great enabling technologies of modern bioinformatics. The transformation of unstructured to structured data using data mapped to ontologies has largely been achieved in a time-consuming manner which relies on human experts. Such manual data curation is accurate but hard to scale and unable to keep pace with the rapid expansion and refinement of both ontologies and the data that we would like to annotate to them. Many bioinformatic tools only support analysis using the Gene Ontology because it is the best annotated and is the most widely used and cited. The delay annotating new ontologies and the lack of support for them in existing analytical tools to aid biological interpretation of data has become a major limitation to their utility and uptake. Thus, I propose that automatic approaches are needed to facilitate the transformation of unstructured data to unlock the potential of all ontologies, with corresponding bioinformatics tools to support their interpretation.*

1.3 Organization of the Thesis

This thesis is divided into three major chapters, two methodological and one application based. In chapter 2, I acknowledge the problem of unstructured gene annotation and the limitation of manual curation. I propose a methodology to ameliorate the problem based on natural language processing techniques (and related implementations) which generates ontology based annotation by mining biomedical text corpora. I discussed the design/implementation details, strengths and weaknesses of the approach and applied the method to generate human gene disease annotation with human disease ontology (HDO). I evaluate the method by validating the generated HDO annotation data with a discussion of extending the annotation of other species. and evaluate/analyse the resulting annotation data.

In chapter 3, I review current enrichment analysis algorithms and tools, and propose an R package that integrates a range of statistical algorithms and topological

methods for ontology based enrichment analysis. Detailed implementation and validation are discussed, and usage of the package is demonstrated by analyzing the activity-regulated cytoskeleton-associated protein (ARC) complex.

In chapter 4, I utilize the data produced with the methodology presented in chapter 2, and the tool discussed in chapter 3 to enrich the understanding of human disease by integrating heterogeneous data in a disease context. I demonstrated the power of the methodologies by 1) building a disease profile for 277 gene classes constructed from three ontologies, and 2) building an disease environment for 1310 human diseases.

A final overall remarks is provided in a concluding chapter in chapter 5, with a discussion of limitations and possible future works.

Chapter 2

Mapping text corpora to ontology terms using natural language processing tools

2.1 NLP in biomedical text mining

Ontology based annotation is a key component in enabling data integration and a wide range of data analysis techniques. However, to annotate genes with a newly developed ontology is time consuming and currently the burden falls to the corresponding ontology consortium. There are numerous text based gene annotation resource freely available from individual labs or from central bioinformatics organization like NCBI or EBI (European Bioinformatics Institute), but it is still a challenge to automate the annotation process with these data. Natural language processing (NLP) has been proved to be a reliable and accurate method to identify and extract relevant data from text corpora. A consecutive series of 18,453 pathology reports were evaluated by Bravo et al. [45] and showed that NLP methods correctly detected 117 out of 118 patients (99.1 %) with prostatic adenocarcinoma after TRUS-guided prostate biopsy.

Concepts (ontology terms), however are difficult to recognize in text due to a disconnect between what is captured in an ontology and how the concepts are expressed in text [130]. Not only the concepts themselves can be expressed in text with a huge amount of variability, ambiguity and underspecification (see above examples), but also the relations among them are vague and rarely described explicitly [131]. Thus, a general approach for concept recognition is still an open research problem.

Ontologies provide a set of terminological resources and semantic constraints,

which are used in a wide range of dictionary-based concept recognition tools. There are recognizers for specific ontologies which are used to address specific categories of terms such as genes or gene products [132], protein mutations [133] or diseases [134, 135], but these recognizers require targeted training material and cannot generically be applied to recognize arbitrary terms from text. Two of the most widely used dictionary-based generic tools for biomedical text mining are the MetaMap [136] developed by the National Library of Medicine (NLM) and the NCBO Annotator [137] developed by the National Center for Biomedical Ontology which are considered the state of the art in the field. Other tools including Whatizit [138], KnowledgeMap [139], CONANN [140], IndexFinder [141], Terminizer [142] and Peregrine [143] are either not freely available or appear not to be in widespread use [144].

MetaMap

The MetaMap [145] is a highly configurable program developed at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. MetaMap can run locally on a Linux machine. It parses input text into noun phrases and generates their variants including alternate spellings, abbreviations, synonyms, inflections and derivations. A candidate set of Metathesaurus concepts were identified and scored based on the strength of mapping from the variants to each candidate concept. MetaMap natively works with UMLS Metathesaurus, but can be optionally configured to work on any ontology. The data file builder (DFB) [146] provided by MetaMap allows the transformation of an ontology into UMLS database tables which is the default dictionary format used by MetaMap.

NCBO Annotator

The NCBO Annotator (formerly referred to as the Open Biomedical Annotator (OBA)), is an ontology-based Web service [52] that annotates textual meta data with biomedical ontology concepts (terms). It allows users to tag their data automatically with ontology concepts. These concepts come from National Center for Biomedical Ontology (NCBO) BioPortal [22], an ontology repository containing more than 500 ontologies (September 2016). The input text is fed into a concept recognition tool developed by the University of Michigan called *MGREP* and ontology annotations are produced only when finding an exact-match. Despite a RESTFUL web service, NCBO also provides a virtual machine which contains a pre-installed, pre-configured version of the

NCBO Annotator that can be run locally on a Linux operating system. It simulates an environment which provides all the pre-requirements (scripts, libs etc.) for the NCBO Annotator and provides the same service locally to the user with a shorter response time. Ontologies that are currently not in the NCBO BioPortal repository can also be added locally.

2.1.1 Comparison of concept recognition tools

Funk et al. [144] carried out a detailed evaluation of the performance of MetaMap, NCBO Annotator and ConceptMapper (*CoM*) [147], a tool dictionary-base recognizer that was not specifically developed for biomedical term recognition. The Colorado Richly Annotated Full-Text (CRAFT) Corpus [148] was used, containing 67 documents (articles) fully annotated with nine biomedical ontologies and terminologies: the Cell Type Ontology, the Chemical Entities of Biological Interest ontology, the NCBI Taxonomy, the Protein Ontology, the Sequence Ontology, the entries of the Entrez Gene database, and the three sub ontologies of the Gene Ontology. The study evaluated the tree recognizers using eight out of the nine ontologies (excluding the Entrez Gene database) in terms of precision and recall under different parameter settings and showed that the best concept recognizers varied from ontology to ontology. They concluded that the generic ConceptMapper, even though not developed for use in the biomedical domain, generally provided the best performance. MetaMap tends to produce the highest recall (five out of eight ontologies tested) but its precision suffers because it finds the most errors; while NCBO annotator produces the highest precision (four out of eight ontologies tested) but falls behind in the recall because it is unable to recognize plurals or variants of terms. As a summarize, the best performance for all tools on all ontologies tested in [144] are shown in fig. 2.1. Besides performance, the study also provided general suggestions on the parameter setting to optimize the performance of the recognizers.

The tool underlying NCBO Annotator, *MGREP*, has been directly compared against the MetaMap on several term recognition tasks [149–151]. These studies reach a similar conclusion as those in [144] that *MGREP* outperforms MetaMap in terms of precision of matching while MetaMap produced more annotations. The actual recall rate was not given in either study because the test corpora used were not fully annotated. This is because MetaMap generates lexical variants on ontology terms during its preprocessing of the ontology while *MGREP* uses exact match. For the same reason,

MetaMap suffers from slow speed which makes it unsuitable for many real-time applications or for applications in which either the data sources or the dictionary changes frequently. On the other hand, *MGREP* has extremely fast execution speed, which makes it possible to process large datasets, that require large dictionaries, or that involve frequent reprocessing.

Many previous studies have been done using NLP tools in the biomedical domain. In terms of disease, MeSH terms [152–155] or OMIM [156, 157] were used for mining human diseases. Rappaport et. al. used an in house process to integrate 44 disease sources into 16919 disease entries named 'disease cards' [46]. These 'disease card' were substantially used in annotating genes based on the information provided by 'gene cards' [158], a similar concept as the 'disease cards' that was used to represent genes and the relevant genomic related information. More recent studies have begun to use the Human Disease Ontology (HDO). Osborne et al. [3] used MetaMap and mapped human disease to HDO terms from the NCBI GeneRIF database, which is a curated database linking genes to short functional annotations and the corresponding publications in the PubMed database. A truth table of the Homayouni gene collection [51] was constructed manually using GeneRIF and OMIM texts as sources which were then used as validation data to evaluate the disease annotation to these two databases. The result reported a 91% recall rate and 97% precision rate of disease annotation using GeneRIF, in contrast with a 22% recall and 98% precision using OMIM suggesting that GeneRIF is a great source for mining gene disease association. LePendu et al. [4] instead used the NCBO Annotator to automatically generate human disease annotation from already existing GO annotations, based on the hypothesis that if a disease term is mentioned in the abstract of the article based on which a GO annotation is created for a gene product, then that disease term is likely to be associated with that gene product. The resulting gene disease annotation dataset was not directly assessed quantitatively but qualitatively with several domain experts by 1) inspecting the recapitulation of known disease associations on well studied genes and 2) examining disease enrichment analysis results for a set of known aging related genes using the generated disease annotation data.

Due to the lack of disease annotation in the CRAFT corpus, it cannot be used to evaluate the performance of the concept recognition tools in terms of finding disease names. In order to assess such feature for MetaMap, NCBO Annotator, and ConceptMapper, an HDO annotated corpus is needed, which is not obviously available. However, such corpus can be generated indirectly, often based on other manually cu-

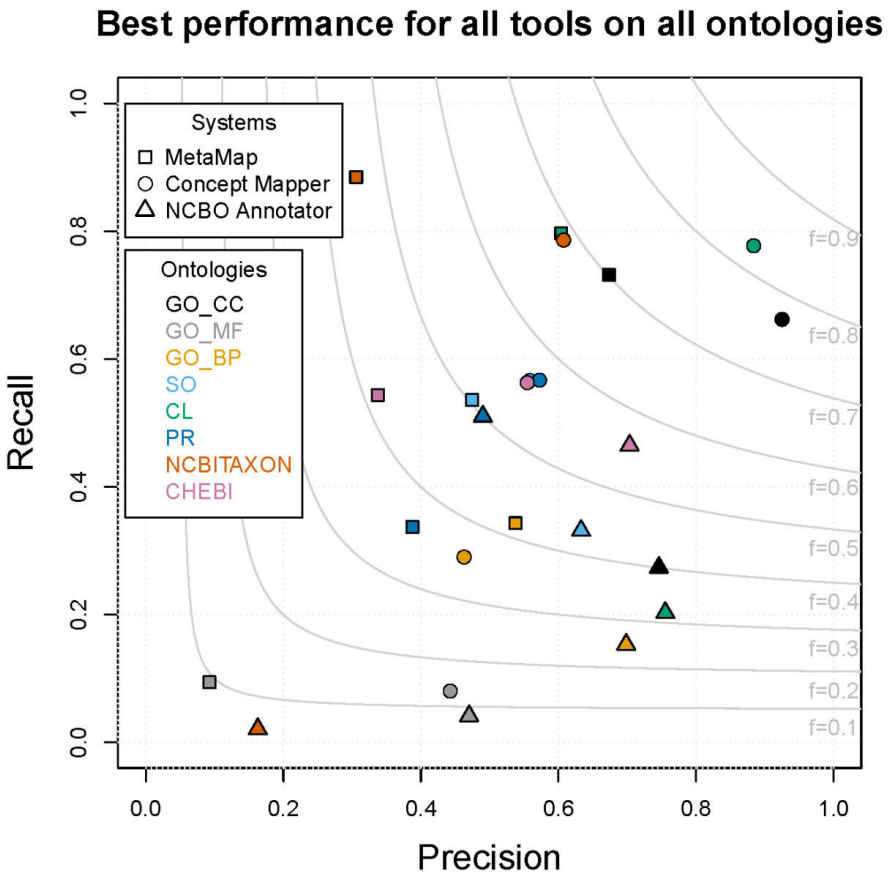


Figure 2.1: Maximum F-measure for concept recognizer-ontology pair. A wide range of maximum scores is seen for each concept recognizer within each ontology [144].

rated resources via annotation cross references. One of such resources is the NHGRI-EBI GWAS Catalog [159], which contains manually curated genome-wide association studies annotated with the Experimental Factor Ontology (EFO) [160]. EFO is a superset of the HDO, making the GWAS Catalog a good resource for generating high-quality HDO annotated corpus.

GWAS Catalog data was downloaded on 28 July 2017, which contained 3050 unique publications annotated with 1576 unique EFO terms, out of which, 304 can be mapped to HDO via EFO-HDO cross references using the EBI Ontology Xref Service [161]. This result in 1548 GWAS publications annotated with HDO terms, from which two HDO annotated corpora (gold standard annotation) are derived: 1) a corpus containing 1535 publication abstracts (13 publication does not have abstract), referred to as *GWAS_Abstract*; and 2) a corpus containing gene rifs(Gene Reference Into Function) from 522 publications from the NCBI GeneRIF database, referred to as *GWAS_GeneRIF*.

The relatively small amount of mapped HDO from EFO is because that EFO covers domains that are not captured in the HDO, i.e, EFO contains not just disease terms, but also terms used to represent other experience factors, for example measurement terms such as 'EFO_0004467 insulin measurement' or biomarkers such as 'EFO_0006842 diabetes mellitus biomarker'. These EFO terms were thus removed from the annotation.

In order to evaluate the performance of MetaMap(*MeM*), NCBO Annotator(*NcA*) and ConceptMapper(*CoM*), HDO terms was used as dictionaries to mine *GWAS_Abstract* and *GWAS_GeneRIF*. The resulting annotation was compared to the gold standard. All comparisons were performed using three comparators: 1) a STRICT comparator (SC), which means that ontology terms generated by each concept recognizer must match the gold-standard annotation exactly to be counted correct; 2) a HIERARCHICAL descendants comparator (HDC), which on top of the STRICT comparator, also counts correct if the ontology terms generated by each concept recognizer are the hierarchical descendants of the gold-standard annotation(see section 1.1.1.4 on page 6 for details of ontology structure); and 3) a HIERARCHICAL comparator (HC) which on top of the HDC, also counts correct for any ancestors of the gold standard. As an example, for the gold standard annotation 'DOID:9351 diabetes mellitus', both 'DOID:4194 glucose metabolism disease'(ancestor) and 'DOID:9744 type 1 diabetes mellitus'(descendant) will be counted correct by HC. Only the latter one will be counted correct by HDC while both will be counted incorrect when using SC.

The performance of the three concept recognizers were measured in terms of true positives (TP), false positives (FP), and false negatives (FN) as well as precision (P), recall (R), and F-measure (F) (see Equation 2.1), which are calculated over all annotations across all corpora in *GWAS_Abstract* and *GWAS_GeneRIF*, i.e. as a *micro-average*(see Equation 2.2).

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F = 2 * \frac{P*R}{P+R} \tag{2.1}$$

$$P_{micro} = \frac{\sum TP}{\sum TP + \sum FP}, R_{micro} = \frac{\sum TP}{\sum TP + \sum FN}, F_{micro} = 2 * \frac{P_{micro} * R_{micro}}{P_{micro} + R_{micro}} \tag{2.2}$$

The evaluation result for each concept recognizers-corpus pair with HDO is presented in fig. 2.2 and table 2.1. The three concept recognizers, within each comparator, have very similar performance in finding HDO terms from the two corpora tested. The best F scores are from the HC comparator, ranging from 0.76 to 0.82 which the worse F scores are from the SC, ranging from 0.42 to 0.52. In general, similarly to the result from Funk et al. [144], *CoM* slightly out performed the other two, but the difference between F scores is subtle, ranging from 0.004 (found between *MeM* and *CoM* with *GWAS_GeneRIF* SC) to 0.065 (found between *CoM* and *NcA* with *GWAS_GeneRIF* HDC).

Interestingly, *NcA* received the lowest F score among the three when using SC and HDC, but the highest score when using HC, indicating that *NcA* often finds HDO terms from the corpora that are the ancestor terms(less specific) of the gold standard. This is partly due to the characteristics of the HDO terms where disease names often consist other less specific disease names, such as 'breast cancer' and 'cancer'. Recognizing 'cancer' is partially correct but such finding of less specific term is not desired. Thus, the HC is likely to be over optimistic, therefore, the result form HD should be consider with caution.

It was also observed that the performance of the concept recognizers vary between corpus. Receiving similar F scores, a higher precision rate was observed for each concept recognizers when annotating *GWAS_GeneRIF*, which suggests that using semi-curated corpus is likely to increase the performance over pure text base corpora such as PubMed abstracts or full text.

In this thesis I use both MetaMap and the NCBO Annotator for NLP analysis of text corpora and integrate them into the same framework to automate the ontology annotation process. Since the two tools have their own advantages and disadvantages, the re-

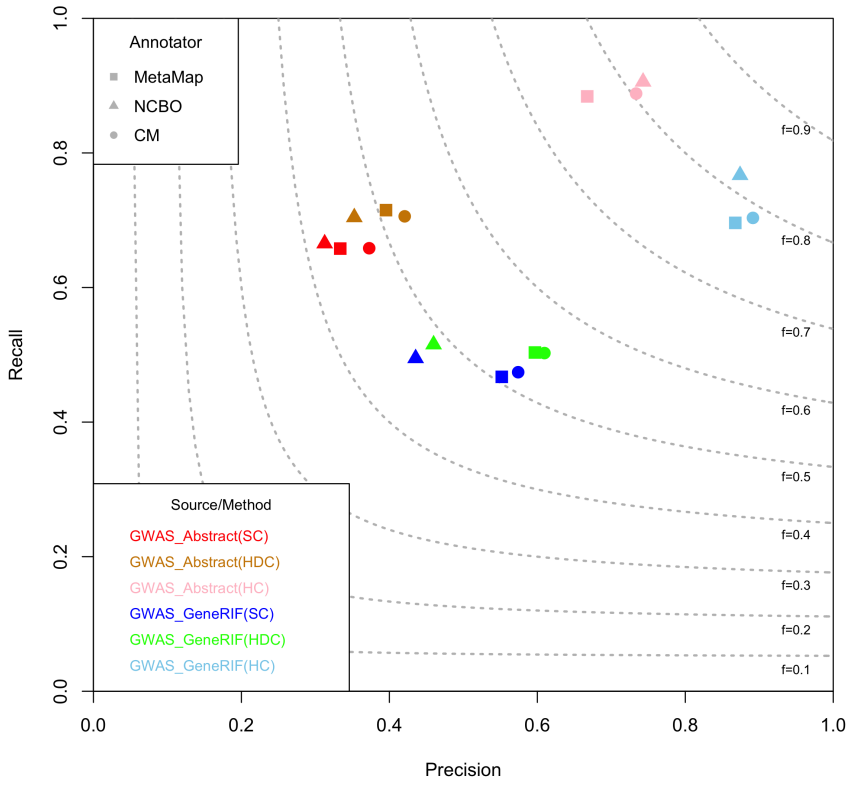


Figure 2.2: Comparison of MetaMap, NCBO Annotator and Concept Mapper in the task of finding Human Disease Ontology terms from publication abstracts and GeneRIFs. Three type of comparators were used to assess the performance, STRICT comparator (SC), HIERARCHICAL descendants comparator (HDC) and HIERARCHICAL comparator (HC). The two HIERARCHICAL comparator correct the result base on the ontology hierarchical structure, thus generated a higher F score than the STRICT comparator.

		MetaMap			NCBO Annotator			Concept Mapper		
		P	R	F	P	R	F	P	R	F
GWAS_Abstract	SC	0.3336	0.6578	0.4427	0.3125	0.6654	0.4253	0.3727	0.6584	0.4760
	HDC	0.3954	0.7149	0.5092	0.3525	0.7043	0.4699	0.4207	0.7057	0.5271
	HC	0.6674	0.8840	0.7606	0.7429	0.9057	0.8163	0.7335	0.8882	0.8035
GWAS_GeneRIF	SC	0.5519	0.4672	0.5060	0.4355	0.4948	0.4632	0.5741	0.4741	0.5193
	HDC	0.5967	0.5034	0.5461	0.4597	0.5153	0.4859	0.6096	0.5025	0.5509
	HC	0.8676	0.6960	0.7724	0.8740	0.7669	0.8170	0.8914	0.7034	0.7863

Table 2.1: Precision, Recall and F score of the result from MetaMap, NCBO Annotator and Concept Mapper on GWAS Catalog corpus with HDO

sulting ontology based annotations can be post-processed based on ontology properties to increase performance. The reason to include MetaMap and the NCBO Annotator in the initial implementation of the framework is that a) they are the most widely used generic (not limited to certain ontology) concept recognizers for biomedical text mining tasks [144] and b) they are being actively developed and are well maintained. Concept Mapper was potentially another powerful concept recognizer, which performed slightly better than the *MeM* and *NcA* (A different in the f score ranging from 0.004 to 0.065 across the corpora tested), but not included in the initial implementation considering that there is no available documentation and the tool has not been updated since 24 Aug 2011 [1]. It, however, will be added to the framework in the next version.

2.1.2 Organization of the chapter

This chapter introduces *OntoSuite-Miner*, a framework composed of a set of Linux shell scripts, Perl/R scripts and two concept recognizers, the MetaMap and the NCBO Annotator, working together to automate the creation of ontology based annotation from publicly available data repositories. What differentiates my method from other approaches is that 1) I use two of the most popular concept recognizer in biomedical text mining with the potential of adding extra recognizers, for example the ConceptMapper, into the *OntoSuite-Miner* framework without too much effort and 2) I reused publicly available curated gene annotation databases as a basis to provide reliable data quality. To demonstrate the feasibility of our method, I generated a Human Disease Ontology (HDO) annotation from three publicly available gene annotation databases including OMIM, GeneRIF and Ensembl variation. HDO is used because 1) disease information is a very interesting and important aspect of gene annotation, but the HDO consortium does not provide gene annotations and 2) the HDO is a good candidate ontology for automatic annotation because disease terms are frequently mentioned in the biomedical text copora (46% more often than GO terms in MEDLINE abstracts [4]), thus the automated annotation process in theory could work better for annotating genes with disease ontology terms than for performing automatic annotation on other ontologies like GO.

In the following sections I provide implementation details of *OntoSuite-Miner*, a discussion of the main design decisions and a number of validation approaches for the generated ontology base annotations. The main output of this chapter are: 1) a general

[1] https://mvnrepository.com/artifact/org.apache.uima/ConceptMapper/2.3.1

framework, which uses unstructured semi-curated gene annotations as a starting point for creating ontology based annotation and 2) a human disease annotation dataset using HDO which enables a wide range of computational analyses and graph/network based analyses for human disease such as disease enrichment analysis.

2.2 Implementation of *OntoSuite-Miner*

2.2.1 Overview

The *OntoSuite-Miner* toolkit is composed of a set of linux shell scripts, perl/R scripts and two text annotators, MetaMap and NCBO Annotator. A high-level schematic describing the implementation of the *OntoSuite-Miner* is shown in fig. 2.3. The main purpose of the toolkit is to link genes with ontology terms given their free text annotation from a variety of sources.

In order to use *OntoSuite-Miner*, the initial dataset must be preprocessed into a list of EntrezGene IDs with the corresponding annotation text in a tabular form stored in a text file. The toolkit provides two methods to input the text, through an interactive command-line accepting one text at a time and return the results 'on the fly'(a user friendly web interface is also available) and, through a command-line batch request which processes all text in a file. Despite the different input methods, the toolkit uses the same underlying methods to process the input data.

The processes can be time consuming depending on the amount of inputted text. Having obtained the text file, a programming model conceptually similar to MapReduce [162] was implemented to speed up the process. Firstly, a work dispatcher unit receives the text file and splits the file into smaller slices. Each slice is then fed to a worker unit (a thread in a multi-core machine or a machine in a cluster) which initiates the mapping process. The work unit dispatches the received text to the two annotators and starts the mapping processing with a set of predefined parameters and a preselected ontology. Once all the work units finish processing the text, the mapping results are collected, merged and indexed by distinct annotation text. A filter is implemented to post process the mapping results with the aim of removing annotation errors occurring in the previous processes (detail discussed below). Finally, EntrezGene IDs are linked to ontology terms through their annotation text and stored in a SQLite relational database. The details of the implementation are described in the following sections.

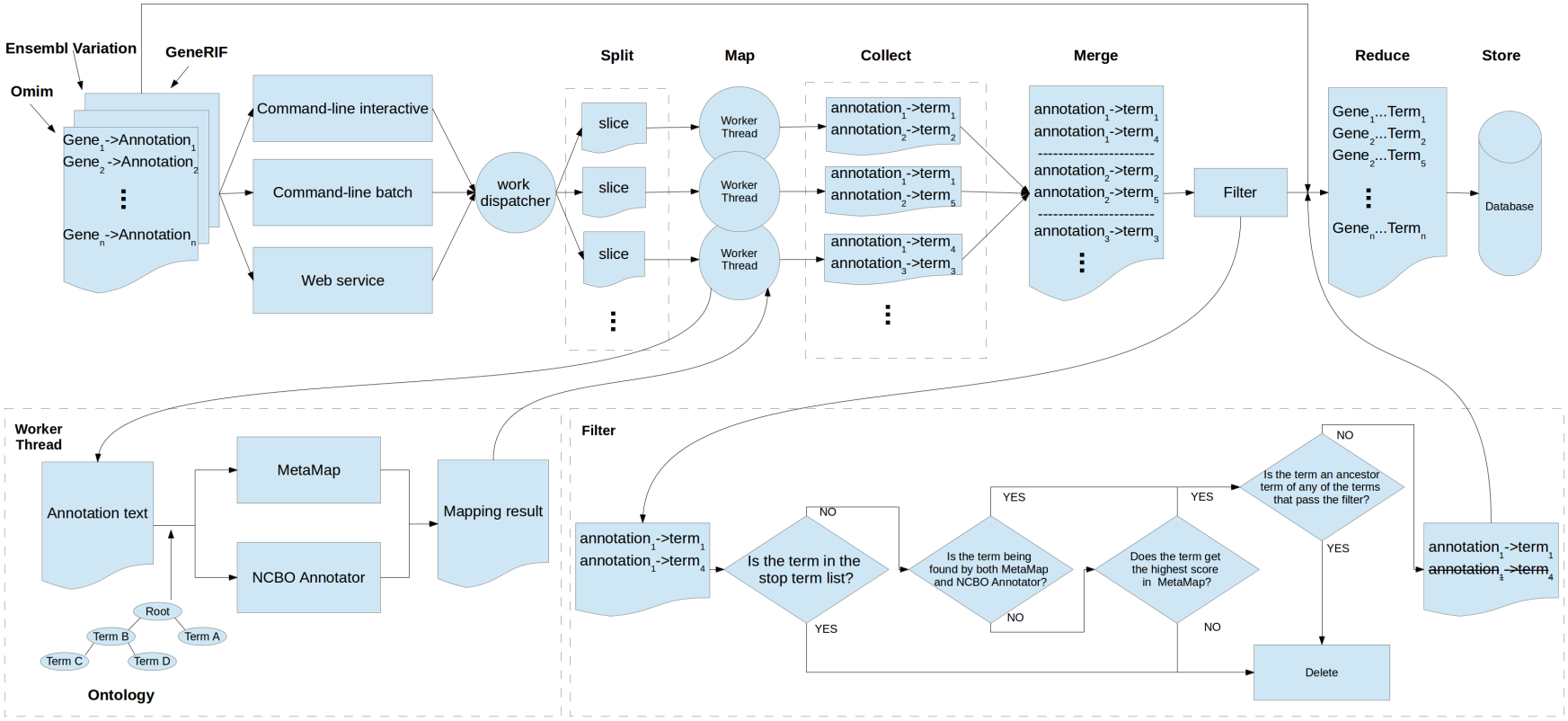


Figure 2.3: A high-level schematic description of the *OntoSuite-Miner* work flow. Text based annotations are parsed and dispatched to worker threads upon receipt. Worker threads perform the ontology based annotations with MetaMap and NCBO-Annotator in parallel and merge the results when all the worker threads finish. This is followed by a filtering process to eliminate errors generated from the two natural language processing tools. The results are then stored in an SQLite relational database.

2.2.2 Description/preprocessing of annotation sources

Annotation source data were taken on 04 Jan 2016 unless otherwise specified. EntrezGene id was used as the primary index for genes.

GeneRIF

GeneRIFs (Gene Reference Into Function) are short annotations of the functions (including diseases) of genes in the NCBI Gene database. A GeneRIF contains a concise phrase describing a function or functions of a gene that exist in the NCBI Gene database. The phrase is restricted to 425 characters in length and requires a published paper in PubMed which describes that function. An example GeneRIF looks like this: *A high expression of CCL19 was a good prognostic factor of lung adenocarcinoma*. GeneRIFs are created by NCBI users who are willing to provide their email address. This Wiki-type resource offers high accuracy and allows a rapid update by the research community [163].

GeneRIF data was downloaded from the NCBI ftp site. Deprecated genes were removed from the data while genes with discontinued id were replaced with the corresponding EntrezGene id according to the discontinued records provided by the NCBI gene history file. There are 369234 distinct geneRIFs referencing 334891 distinct PubMed articles and 16359 distinct genes in the GeneRIF dataset. This represents 86% of the roughly 19,000 protein coding genes estimated to exist in the human genome [164].

OMIM

The Online Mendelian Inheritance in Man (OMIM) database is part of NCBI which is manually curated and contains information on human genes and genetic disorders. Human genes and disorder were indexed with OMIM's own access number, locus_mim_acc and disorder_mim_acc. A typical OMIM entry is as follows: "*602290-615988-Bardet-Biedl syndrome 11*", which indicates that gene with locus_mim_acc 602290 is associated with *Bardet-Biedl syndrome 11* with disorder_mim_acc 615988. OMIM's manual curation process makes its data very precise, and is an excellent source for mining gene disease associations. However, there is a noticeable delay in updating the database and a limited gene/disease coverage because the curation process is extremely time consuming.

OMIM data was taken from the NCBI ftp site. Genes with locus_mim_acc were mapped to EntrezGene id according to the mapping provided in the mim2gene.txt file

of OMIM. Those genes without a mapping to an EntrezGene entry were removed. As a result, 3596 distinct genes were annotated with 4482 disorders in the OMIM dataset. However, subtypes between the same disorder in OMIM are considered as different disease entities. For example, *Bardet-Biedl syndrome 11* and *Bardet-Biedl syndrome 12* in OMIM are considered two disorder entities instead of a single *Bardet-Biedl syndrome* entity. Thus, the number of distinct disorder in OMIM is far less than 4482 and the actual number of distinct disorder is hard to be determined.

Ensembl variation

Genetic recombination is an event naturally occurring during meiosis. It is facilitated by chromosomal crossover where homologous chromosomes tightly pair together and exchange genetic information (fig. 2.4). These exchanges increase genome diversity and are essential for proper chromosome segregation at the first meiotic division. recombination can occur with small probability at any location along chromosome and the frequency of recombination between two locus depends on the distance between them. Therefore, Recombination will rarely separate loci that lie closely together on a chromosome because it is unlikely that a crossover is located in such a small space between the two loci. Thus, sets of alleles on the same small chromosomal segment tend to be transmitted as a block through generations. Therefore, SNPs are likely to be inherited together with their closely located up-stream/down-stream genes during evolution. Most SNPs are found located between genes, which act as good biomarkers of diseases. The other SNPs occur within genes or in regulatory regions near genes which may play a more direct role in disease by affecting directly on the genes' function. Since the completion of the human genome and following various technological advances, an increasing number of GWAS studies identified SNPs which provided evidence for possible disease association of many previously uncharacterized genes [165, 166]. Such data is potentially a very good resource for mining gene disease association.

Ensembl variation [168] stores areas of the genome that differ between individual genomes ("variants") and, where available, their associated disease and phenotype information. Different types of variants exist including SNP, Insertion/Deletion and more complicated structural variants. I am interested in SNP data in this project because SNP is the most dominant type of variants in human (fig. 2.5, 'SNV' is the corresponding term for SNP in Sequence Ontology), accounting for 89% of all types of variants. As the number of GWAS studies is growing rapidly, a lot of new SNPs

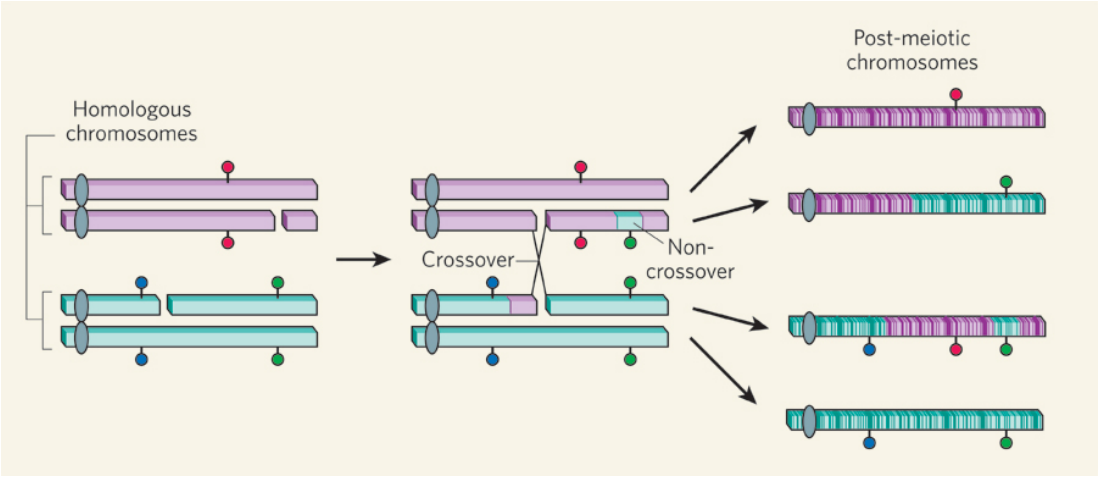


Figure 2.4: Genetic recombination event during meiosis. It is facilitated by chromosomal crossover where homologous chromosomes tightly paired together and exchange genetic information [167].

are being associated with diseases, which provides a great resource for linking these diseases to genes. For example, SNP rs1333049 is located on chromosome 9 and has been found to be associated with coronary heart disease. Gene 'CDKN2B-AS1' and 'RP11-408N14.1' are the closest up- and down-stream genes of the SNP so it can be inferred that they are potentially associated with coronary heart disease.

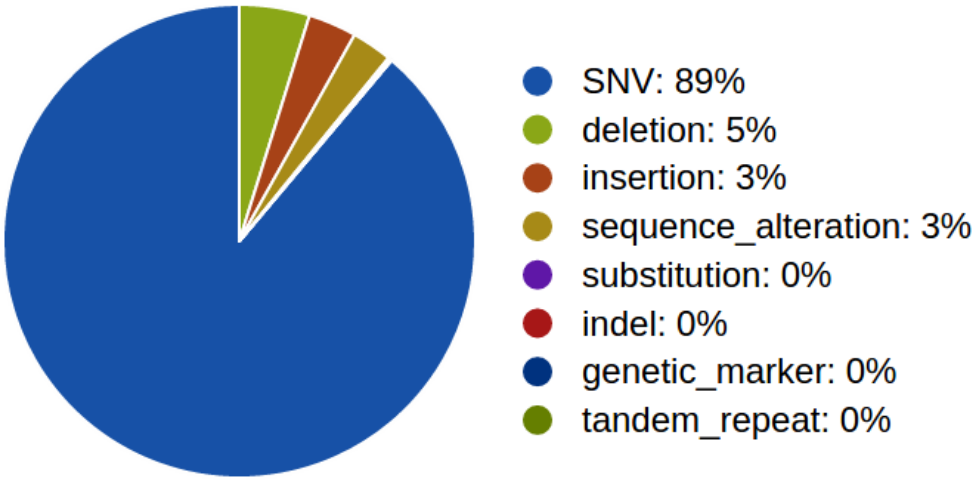


Figure 2.5: Human variant type distribution with Sequence Ontology(SO) terms in the Ensembl variation database. SNV is the corresponding term for SNP in SO.

Not all SNPs have diseases/phenotypes associated with them. The Ensembl variation database uses a concept 'variant sets' to group variants that share some property

together. I used BioMart [169] to search and download all of the human SNPs which are in the variation set 'All phenotype/disease-associated variants', that is all the SNPs that have at least one disease/phenotype associated with them. These SNPs were gathered from 11 different sources (fig. 2.6) including ClinVar [170], NHGRI-EBI GWAS Catalog [165] and OMIM [86]. As shown in table 2.2, the majority of these GWAS studies were cancer studies including breast cancer, ovarian cancer and colorectal cancer. This is because, compared to the traditional candidate gene approach, GWAS provides a powerful approach to identify common disease loci without prior knowledge of gene, location or function, thus are particular useful in the study of polygenic diseases. In total, 74715 unique SNPs were found in the variation database and 17088 unique genes were identified as the most up-stream/down-stream genes or the overlapping genes of the SNPs.

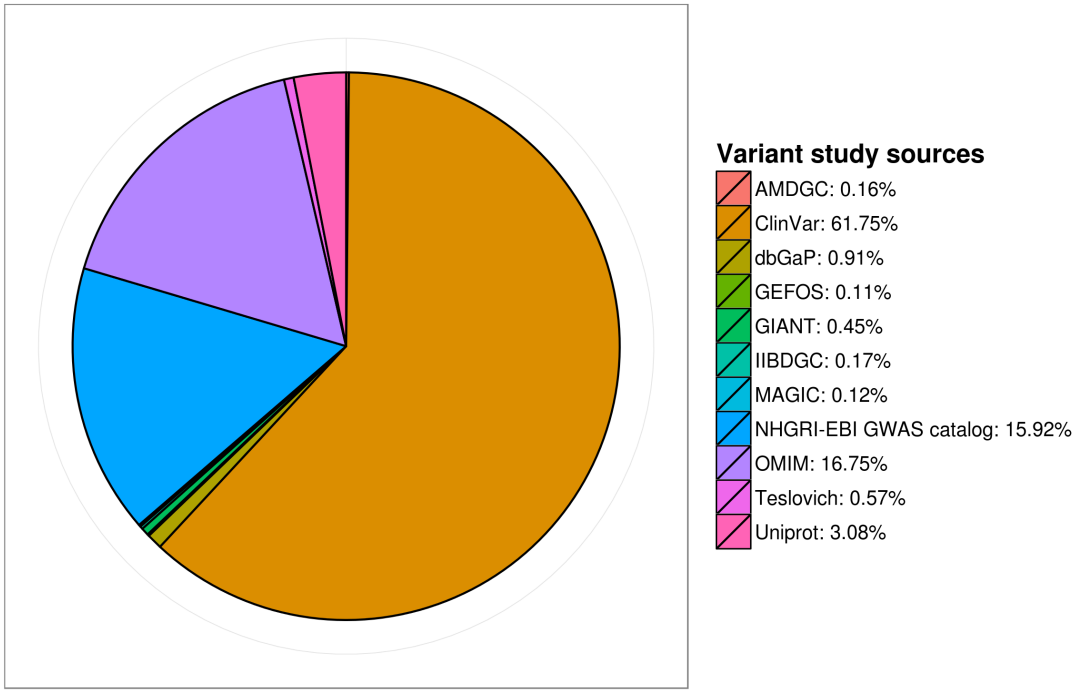


Figure 2.6: Source distribution in the Ensembl variation database. SNPs data were integrated from 11 sources with more than half of the SNPs from the ClinVar database [170].

	Phenotype	count
1	Hereditary cancer-predisposing syndrome	27885
2	Familial cancer of breast	13748
3	Breast-ovarian cancer familial 2	7269
4	Lynch syndrome	6882
5	Cardiomyopathy	6657
6	Tuberous sclerosis syndrome	6072
7	Breast-ovarian cancer familial 1	5535
8	CYSTIC FIBROSIS	3532
9	Height	3372
10	Cardiac arrhythmia	3276
11	Familial colorectal cancer	2970
12	Congenital long QT syndrome	2327
13	Obesity-related traits	2154
14	Thoracic aortic aneurysms and aortic dissections	2058
15	Rasopathy	1705
16	Severe myoclonic epilepsy in infancy	1690
17	Alport syndrome X-linked recessive	1539
18	BRCA1 and BRCA2 Hereditary Breast and Ovarian Cancer	1458
19	Malignant tumor of prostate	1370
20	Primary familial hypertrophic cardiomyopathy	1109

Table 2.2: Top 20 phenotypes annotated in the Ensembl variation database. The majority of the studies were cancer studies including breast cancer, ovarian cancer and colorectal cancer.

The Independence between sources

Even through the three data sources discussed above, namely GeneRIF, OMIM and Ensembl variation, are created and maintained independently by different bioinformatics organizations, there are overlapping information among them. In order to assess the degree of independence between the sources, gene/pubmed pair was used to identify overlapping data entities. GeneRIF and Ensembl variation store pubmed id in their data entities whenever they are available, but OMIM does not contain such data. Thus, the assessment was carried out only between GeneRIF and Ensembl variation.

GeneRIF contains 482084 unique gene/pubmed pairs while the corresponding number in Ensembl variation is 71909, among which 1830 are share between them. Such a small overlapping indicates that the two data sources are reasonably independent. Further assessment is needed to evaluate the dependence of OMIM against the other two, when such data become available from OMIM.

2.2.3 Annotator setup/configuration

Two publicly available dictionary-based annotators were implemented locally on a Linux server, NCBO Annotator [137] and MetaMap [136]. These two annotators are both able to produce annotations for ontologies but differ in their underlying implementation and amount of configurable parameters. Funk et.al [144] evaluated three annotators including NCBO Annotator and MetaMap on eight biomedical ontologies in the Colorado Richly Annotated Full-Text(CRAFT) Corpus [171]. Over 1000 parameter combinations were examined by Funk et.al and best-performing parameters for each ontology were presented. Despite the small difference between ontology-parameter pairs, a general guild-line for choosing parameters was discussed and suggestions for choosing the best parameters based on ontology characteristics were presented. The parameters of NCBO Annotator and MetaMap in *OntoSuite-Miner* were set up based on the best-performing parameters with modification (see below for details) to suit the Human Disease Ontology used in this project.

2.2.3.1 NCBO-Annotator

The NCBO Annotator provides a RESTFUL web service[2], as well as a virtual machine which contains a pre-installed, pre-configured version of the NCBO Annotator running on a Linux operating system. It stimulates an environment which provide all the pre-

[2]https://bioportal.bioontology.org/annotator

requirements(scripts,libs, etc.) for the NCBO Annotator and provides the same service locally to the user. NCBO virtual machine (hereby referred to as *NcA*) version 2.4 was integrated locally in *OntoSuite-Miner*, because the following reasons: 1) A local implementation provides a shorter response time compared to using the RESTFUL web service remotely; 2) The user has full control of the annotator's configurations; 3) having the annotator service locally ensures that the general maintenance on the NCBO Annotator web service does not interrupt *OntoSuite-Miner*; and 4) The RESTFUL web service alway use the most updated ontologies, thus may produce inconsistence mapping result due to the update of ontologies. Having a local service allows the user to precisely control the ontology version for consistent performance, as well as the possibility of using customized ontology that does not exist in the NCBO ontology repository.

In order to keep ontologies consistent within *OntoSuite-Miner*, instead of using the ontologies provided by NCBO directly, ontology data were manually parsed into *NcA* following the instructions described on the NCBO virtual machine wiki[3]. This also enables customized ontologies to be used in *NcA*. In this project, the Human Disease (HDO) and Human Phenotype (HPO) ontologies were downloaded (04-12-2015) and parsed into *NcA*.

NcA is highly configurable and the parameter settings can make a significant difference to its performance. According to [144], a general rule for *NcA* was that only whole words should be matched and synonyms of terms should be used. Thus, parameter *whole_word_only* was set to *true* and parameter *exclude_synonyms* was set to *false*. The Minimum term size was reported not to effect the matching of terms but to act as a filter to remove matches of less than a certain length. A safe value of one or three was suggested to remove only very small words. However, in the case of HDO, diseases sometimes contains synonyms of abbreviation which were frequently used in biomedical text. For example, *AD* for Alzheimer's disease in the sentence *These results suggest that variants of APOA1 might influence the onset and the risk for AD* and *COPD* for chronic obstructive pulmonary disease in the sentence *The CYP2E1 and NAT2 variants associated with COPD*. Such information would be omitted if a minimum term size was set to one or three. Thus, based on the characteristic of HDO, parameter *minimum_match_length* was set to 0. Note that having a 0 minimum match length increases the chance of identifying ontology concepts in a sentence with the price of introducing abbreviation annotation error, which is discussed further in sec-

[3] http://www.bioontology.org/wiki/index.php/Virtual_Appliance_FAQ

tion 2.3.4.1 on page 76. Other parameters that did not impact the performance of *NcA* were set to their default value and a set of default stop words were used and can be found on the NCBO annotator documentation website[4].

2.2.3.2 MetaMap

MetaMap is a highly configurable program designed to work with the UMLS Metathesaurus, but can be optionally configured to work on ontology. The data file builder (DFB)[5] provided by MetaMap allows the transform of UMLS-like database tables into a dictionary format used by MetaMap. Thus, Perl scripts were written in *OntoSuite-Miner* to convert ontology from a standard OBO format to UMLS database tables following the specification in the DFB, which is subsequently used in the MetaMap.

In contrast to a web service like NCBO Annotator, MetaMap runs natively on a Linux machine. MetaMap v.2013 was installed and integrated in *OntoSuite-Miner* and hereby referred to as MeM. The Human Disease (HDO) and human Phenotype (HPO) were downloaded (04-12-2015) and parsed into UMLS database tables using the Perl scripts. DFB was used to load the ontologies into MeM.

Parameters were set according to [144] for best performance. Gaps between words (parameter $-g$) was not allowed when matching. Two data models are available in MeM, the *Strict* model and the *Relaxed* model which differs in the way MeM partitions its dictionary. The *Relaxed* Model is a proper superset of the *Strict* Model, and typically contains dictionary strings containing internal syntactic structure, such as conjunction like 'arms and legs'. The *STRICT* model, on the other hand, performs an extra filtering step on the dictionary terms (ontology terms), which was reported to increase precision without losing recall for two out of eight ontologies tested in [144] (no difference in the other six ontologies). It is also documented to produce the highest level of accuracy and is used as the default model by MeM. Therefore, *Strict* model was used in the current implementation of MeM in *OntoSuite-Miner*. Enabling reordering (parameter $-i$) of words in the terms is another way to configure MetaMap to recognize more complex terms but was discouraged because a drop of precision was observed without an increase in recall [144].

One unique feature of MeM is that it is able to compute acronym and abbreviation variants when mapping text to the dictionary terms (ontology terms). It is observed that the use of all acronym/abbreviations (parameter $-a$) introduced many erroneous

[4] http://data.bioontology.org/documentation

[5] https://metamap.nlm.nih.gov/DataFileBuilder.shtml

matches, resulting in a decrease in precision without an increase in recall [144]. The use of acronym/abbreviation that has unique expressions (parameter $-u$) was thus recommended and configured in *MeM*. Minimum term size, acts exactly like textitminimum_match_length of the *NcA* parameter, was set to 0 due to the same reason discussed early in the parameter setting of *NcA*.

MeM can be configured to generate derivational variants which help to identify different forms of terms, for example, 'cancer-cancerous' as a noun-adjective derivational. The aim of using derivational is to increase recall without introducing ambiguous terms. There are three values (*all*, *none*,and *adjnounonly*) for this parameter and it produces the most varied results. The default value *adjnounonly* was suggested in [144] if most of the ontology terms could be expressed as nouns or verbs, otherwise hurts the performance if ontology terms do not follow traditional English rules, like gene/protein names. Human disease ontology belongs the former, thus generating derivational variants was allowed for *adjnounonly*.

MeM produces a score for each of its mapping result in the range of 0 to 1000, with 1000 being the most confident. A threshold of the score can be used to filter results. A high threshold (only keep high scored mapping) results in a high precision with low recall while a threshold of 0 returns all mappings, resulting in the highest recall with the lowest precision. Performance is best on all ontologies tested in [144] when using most of the mapping found by *MeM*, so a score of 0 is suggested, thus implemented *MeM*. Other parameters that did not impact the performance of *MeM* were set to their default value as documented by MetaMap[6].

2.2.3.3 The worker thread

MeM and *NcA* are coordinated by a work worker thread. After text corpora are loaded into *OntoSuite-Miner*, a work dispatcher split the corpora into a number (10 in the current implementation) of smaller pieces containing part of the text corpora. Each piece is then dispatched to a worker thread (a thread in a multi-core machine or a machine in a cluster) where the annotation process is initiated. With in a worker thread, Linux shell scripts are used to dispatch received text corpora to *MeM* and *NcA* simultaneously and gather the outputs. Worker threads perform the annotation process in parallel and results were merged when all the worker threads finish.

[6]https://metamap.nlm.nih.gov/Docs/MetaMap13_Usage.shtml

2.2.4 The filtering process

As shown in fig. 2.3, an extra filtering step was implemented to post-process the merged annotation. The filtering process was not designed to discover more annotations but to remove unwanted annotations and possible errors. For a particular annotation, firstly, the filter checks if the annotated term was present in a predefined list of unwanted terms. The unwanted term list for HDO contains two terms, 'DOID:4 disease' and 'DOID:225 syndrome'. Secondly, the annotation was filtered based on the annotator used to produce the annotation. If the annotation was found by both *MeM* and *NcA*, it would be kept under the rationale that an annotation agreed by both annotators was more reliable. MetaMap computed a score for each annotation based on the strength of the mapping, thus the annotation was kept if it was found solely by *MeM* and was the top ranked annotation. *NcA* did not provide any score for its result, and it is generally more accurate then *MeM*, thus all annotations from *NcA* were kept at this stage. Finally, the annotation was examined, together with other annotations that passed the filter previously, for their relationship within the ontology structure hierarchy. Ancestor terms (more generic terms) were removed if there was an offspring term (more specific terms) present, ensuring that only the most specific (informative) terms were kept. For the Human Disease Ontology it is very common that some disease names partly compose other disease names. Some of these diseases have a parent-child relation in the ontology hierarchy, for example, 'DOID:162 cancer' and 'DOID:1612 breast cancer' which is a type of cancer. However, this is not always the case, for example, 'DOID:1091 tooth disease' which is a dental disorder and a completely different disease 'DOID:10595 Charcot-Marie-Tooth disease' which is a neuromuscular disease. Such characteristics of disease names usually confuses the annotators, leading to an annotation error called *coordinating conjunctions* (Detail discussed in section 2.3.4.1 on page 76. This final filtering step, accounting for the ontology structure, was therefore implemented to deal with such errors. An example of the filter process was shown in fig. 2.7.

2.2.5 Data storage

OntoSuite-Miner use portable SQLite database to store the final annotation as well as all the intermediate results that are needed to trace back any annotation to its source. R scripts implemented in the *OntoSuite-Miner* allow a user to pull all of the evidence associated with a particular gene disease association (fig. 2.8). A gene disease asso-

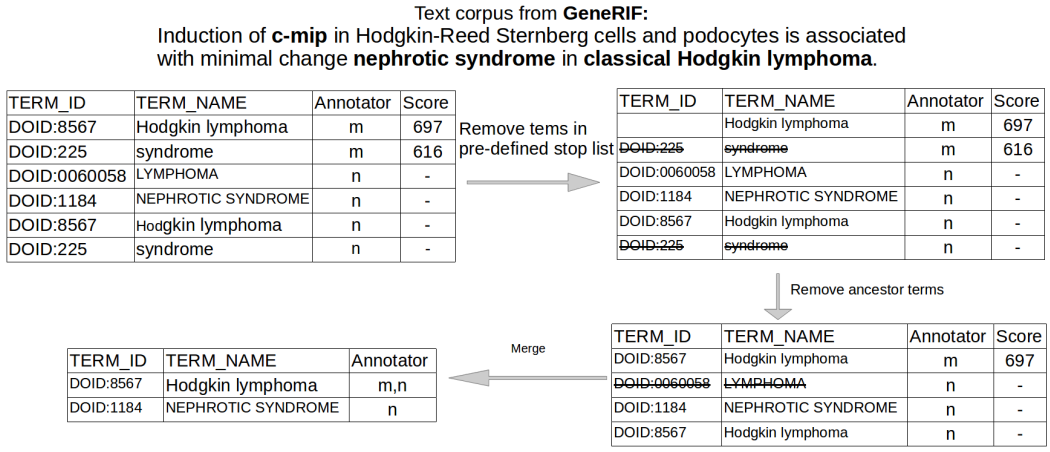


Figure 2.7: An example of the filtering process. A GeneRIF was fed into *OntoSuite-Miner* and 2 HDO terms were found by MeM and 4 HDO terms by NcA. 'DOID:225 syndrome' was a predefined unwanted term and was removed. 'DOID:0060058 LYMPHOMA' was removed because it is a parent term of 'DOID:8567 Hodgkin lymphoma'. As a result, two HDO terms were annotated to this GeneRIF, 'DOID:8567 Hodgkin lymphoma' and 'DOID:1184 NEPHROTIC SYNDROME'.

ciation file was pre-computed that provides only the gene disease association without evidence for easy access.

```
          7157       DOID:1612           score
         "TP53" "breast cancer"         "0.717"
##############Overview
  entrez_id   term_id source mappting_tool evidence
1      7157 DOID:1612      g             m       26
2      7157 DOID:1612  o,g,v           m,n      192
3      7157 DOID:1612      g             n        5
##############OMIM
  entrez_id   term_id     term_name locus_mim_acc disorder_mim_acc     disorders           gene_symbols mapping_tool
1      7157 DOID:1612 breast cancer        191170           114480 Breast cancer TP53, P53, LFS1 , BCC7          m,n
##############GENERIF
  entrez_id   term_id     term_name       pubmed_id                                                    rif mapping_tool
1      7157 DOID:1612 breast cancer pubmed/11786482 These results demonstrate that butyrate inhibited ...          m,n
2      7157 DOID:1612 breast cancer pubmed/11788906 Telomerase activity in microdissected human breast...          m,n
3      7157 DOID:1612 breast cancer pubmed/11852106 Curcumin induces apoptosis in human breast cancer ...          m,n
4      7157 DOID:1612 breast cancer pubmed/11872638 We investigated three common sequence variants in ...          m,n
5      7157 DOID:1612 breast cancer pubmed/11883440 minor role of exon 5-9 among Sudanese breast cance...          m,n
6      7157 DOID:1612 breast cancer pubmed/11953857 p53 mutational pathway may favor selection for Erb...          m,n
7      7157 DOID:1612  breast tumor pubmed/12101184 role in adriamycin-induced senescence in breast tu...          m,n
8      7157 DOID:1612 breast cancer pubmed/12209590 TP53 mutations in breast cancer tumors of patients...          m,n
...
##############ENSEMBL VARIATION
  entrez_id   term_id     term_name variation_id relative_position phenotype_description mapping_tool
1      7157 DOID:1612 breast cancer   rs28934874                 o BREAST CANCER SOMATIC          m,n
```

Figure 2.8: Supporting evidence for linking gene ‘TP53’ to HDO term ‘DOID:1612 breast cancer’ from OMIM, GeneRIF and Ensembl variation.

2.3 Application of *OntoSuite-Miner*

2.3.1 Evaluating the performance of *OntoSuite-Miner*

In order to demonstrate the performance improvements of *OntoSuite-Miner*, over MetaMap or NCBO Annotator alone, they are used to annotate the *GWAS_Abstract* and the *GWAS_GeneRIF* corpus (see section 2.1.1 for a description of the corpus and comparison methods) corpus with Human Disease Ontology. The results were three HDO annotation data set generated from *MeM*, *NcA* and *OntoSuite-Miner* respectively. Because *OntoSuite-Miner* implemented multiple concept recognizers, a high confident subset, supported by all implemented concept recognizers in *OntoSuite-Miner* (referred to as *OntoSuite-Miner_MN*) were extracted from the *OntoSuite-Miner* result and included in the substantial evaluation.

The above four HDO annotation data sets were assessed in terms of precision, recall and F scores (see eq. (2.2) on page 39) using the STRICT Comparator (SC) and the HIERARCHICAL Descendants Comparator (HDC). The HIERARCHICAL Comparator was overly optimistic, thus not included here. The detail comparison methods were described previously in section 2.1.1. The result was shown in fig. 2.9 and table 2.3.

The best F score (0.6153) was generated from *OntoSuite-Miner_MN* from mining *GWAS_Abstract* corpus with the HIERARCHICAL Descendants Comparator (HDC) while the best precision rate was reported in also from *OntoSuite-Miner_MN* from mining *GWAS_Abstract* with the HDC, where 78% of the mapping are correct. *OntoSuite-Miner_MN* was able to recover 72% of the annotation from *GWAS_Abstract* but around 50% of the mapping generated were false positives. An improved F score was observed in all cases from *OntoSuite-Miner* (also its high confident subset) comparing to using *MeM* or *NcA* along, suggesting the an increasing performance from the integration of multiple concept recognizers.

		MetaMap			NCBO Annotator			OntoSuite-Miner			OntoSuite-Miner_MN		
		P	R	F	P	R	F	P	R	F	P	R	F
GWAS Abstract	SC	0.3336	0.6578	0.4427	0.3125	0.6654	0.4253	0.3969	0.6233	0.4849	0.5621	0.5488	0.5554
	HDC	0.3954	0.7149	0.5092	0.3525	0.7043	0.4699	0.4752	0.7271	0.5748	0.6252	0.6057	0.6153
GWAS GeneRIF	SC	0.5519	0.4672	0.5060	0.4355	0.4948	0.4632	0.6452	0.5017	0.5645	0.7387	0.4241	0.5388
	HDC	0.5967	0.5034	0.5461	0.4597	0.5153	0.4859	0.6984	0.5403	0.6092	0.7807	0.4482	0.5695

Table 2.3: Precision, Recall and F score of the result from *OntoSuite-Miner*, MetaMap and NCBO Annotator on GWAS Catalog corpus with HDO

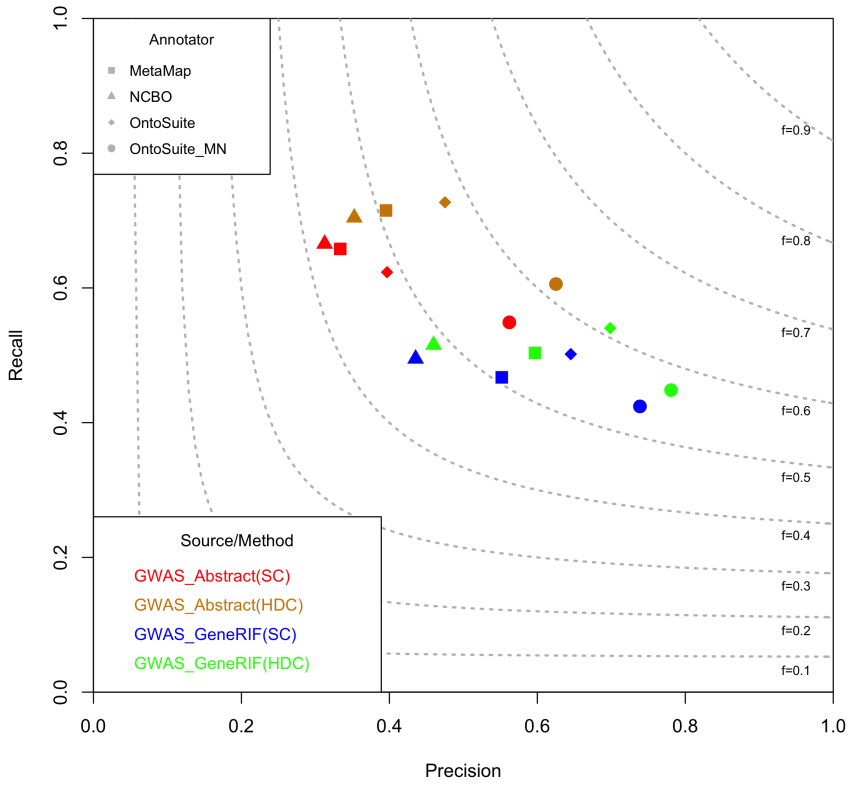


Figure 2.9: Comparison of the performance of *OntoSuite-Miner*, MetaMap and NCBO Annotator in the task of finding Human Disease Ontology terms from testing corpus of publication abstracts and GeneRIFs generated from the GWAS Catalog. Two type of comparators, a STRICT comparator (SC) and a HIERARCHICAL descendants comparator (HDC)(see section 2.1.1, were used to assess the performance in terms of Precision, Recall and F score. A high confident subset of the annotation generated by *OntoSuite-Miner*, referred to as *OntoSuite-Miner_MN* is also assessed. An overall higher F score was observed suggesting from *OntoSuite-Miner*, suggesting the improvement of the result by integrating multiple concept recognizers.).

2.3.2 Creating the Human Disease Gene Database (*HDGDB*)

I applied *OntoSuite-Miner* on three publicly available data sources described early including OMIM, GeneRIF and Ensembl variation, to find HDO terms in the gene annotation text corpora. As a result, three HDO based gene annotation datasets were produced, one for each of the three sources. The transformation of the unstructured, text based annotation to HDO base annotation allows easy comparison and integration of the annotation. Therefore, a comprehensive disease annotation database named Human Disease Gene Database (*HDGDB*) was created, by merging the three HDO annotations datasets. EntrezGene id was used as the primary index of the annotation. In the following sections, I will report the disease annotation from each source, and from *HDGDB*.

OMIM

There are 5267 distinct OMIM disease entities referencing 3596 distinct genes in the OMIM database. *OntoSuite-Miner* was able to find 1248 unique HDO term in 3971 (75.39%) OMIM disease entities, annotating 2914 distinct human genes with at least one HDO term, generating in total 4759 unique gene disease associations. On average, 1.18 HDO terms were found for each OMIM disease entity. The top 20 most annotated genes and diseases in OMIM are shown in table 2.4.

GeneRIF

There are 369102 distinct geneRIFs referencing 334789 distinct PubMed articles and 16103 genes in the NCBI GeneRIF database (data taken on 16 Feb 2015). *OntoSuite-Miner* was able to find 3303 unique HDO term in 154202 (41.78%) rifs, annotating 10977 distinct human genes with at least one HDO term, generating in total 111875 unique gene disease associations. On average, 1.25 HDO terms were found for each RIF of the mapped RIFs. The top 20 most annotated genes and diseases in the GeneRIF database are shown in table 2.5.

Ensembl variation

There are 10707 distinct disease entities referencing 17085 distinct genes in the Ensembl variation database. *OntoSuite-Miner* was able to find 1421 unique HDO term in 6267 (58.53%) disease entities, annotating 13853 distinct human genes with at least one HDO term, generating in total 39993 unique gene disease associations. On average, 1.19 HDO terms were found for each disease entity. The top 20 most annotated

	Gene	Count
1	FGFR2	12
2	TP53	12
3	PTEN	11
4	FGFR3	10
5	LMNA	10
6	COL2A1	9
7	KRAS	9
8	BRAF	8
9	CDH1	8
10	GJA1	8
11	PIK3CA	8
12	PSEN1	8
13	VCP	8
14	GDF5	8
15	CHCHD10	8
16	ATM	7
17	BRCA2	7
18	TRPV4	7
19	ABCA4	6
20	CLCN5	6

	HDO	Count
1	intellectual disability	81
2	retinitis	68
3	retinitis pigmentosa	67
4	autosomal recessive nonsyndromic deafness	54
5	cataract	49
6	Charcot-Marie-Tooth disease	46
7	tooth disease	46
8	hereditary spastic paraplegia	46
9	autosomal dominant cerebellar ataxia	42
10	carcinoma	41
11	congenital disorder of glycosylation	38
12	autosomal dominant non-syndromic intell...	36
13	microcephaly	36
14	dilated cardiomyopathy	36
15	myopathy	34
16	Alzheimer's disease	33
17	colorectal cancer	33
18	autosomal dominant nonsyndromic deafness	31
19	Ohtahara syndrome	31
20	schizophrenia	29

Table 2.4: HDO annotation produced by *OntoSuite-Miner* from the OMIM database. The top 20 most annotated genes in the table on the left and the top 20 most annotated diseases in the table on the right. The 'Count' is the number of unique gene/disease a disease/gene is annotated with. 'Intellectual disability' received the most research attention together with diseases like 'Retinitis pigmentosa', reflecting the fact that OMIM focuses on studies of Mendelian disorders.

genes and diseases in the Ensembl variation are shown in table 2.6.

	Gene	Count
1	TNF	426
2	VEGFA	413
3	TP53	398
4	IL6	371
5	TGFB1	347
6	MMP9	313
7	IL10	270
8	EGFR	252
9	CXCL8	249
10	MTHFR	246
11	CDKN2A	238
12	IL1B	235
13	PTGS2	233
14	MMP2	224
15	TLR4	219
16	NFKB1	208
17	CRP	207
18	HIF1A	205
19	STAT3	200
20	CCL2	199

	HDO	Count
1	cancer	3140
2	carcinoma	2973
3	breast cancer	2815
4	hepatocellular carcinoma	1919
5	prostate cancer	1789
6	lung cancer	1742
7	colorectal cancer	1587
8	squamous cell carcinoma	1442
9	adenocarcinoma	1310
10	ovarian cancer	1186
11	melanoma	1172
12	malignant glioma	1024
13	Alzheimer's disease	1011
14	pancreatic cancer	999
15	schizophrenia	933
16	colon cancer	927
17	hepatitis	822
18	non-small cell lung carcinoma	761
19	rheumatoid arthritis	759
20	esophageal carcinoma	667

Table 2.5: HDO annotation produced by *OntoSuite-Miner* from the GeneRIF database. The top 20 most annotated genes in the table on the left and the top 20 most annotated diseases in the table on the right. The 'Count' is the number of unique gene/disease a disease/gene is annotated with. Various cancers like 'breast cancer' and 'lung cancer' and their corresponding genes were found frequently in the database. Other diseases like 'Alzheimer's disease' and 'schizophrenia' also attracted a lot of research attention.

	Gene	Count
1	HLA-DQB1	44
2	HLA-DQA1	38
3	LOC102725019	37
4	HLA-DRB1	36
5	C6orf10	34
6	HLA-DRA	33
7	NOTCH4	30
8	LOC100287015	29
9	TP53	28
10	ATP1B2	27
11	CSMD1	27
12	WRAP53	26
13	HCG23	26
14	LOC101927815	25
15	SEMA4A	24
16	CDKN2B-AS1	24
17	HLA-B	23
18	LMNA	23
19	BTNL2	23
20	MIR4457	23

	HDO	Count
1	obesity	2176
2	cancer	1452
3	schizophrenia	999
4	bipolar disorder	911
5	Alzheimer's disease	557
6	attention deficit hyperactivity disorder	555
7	breast cancer	519
8	lateral sclerosis	489
9	major depressive disorder	488
10	prostate cancer	478
11	intellectual disability	457
12	rheumatoid arthritis	457
13	Usher syndrome	448
14	multiple sclerosis	437
15	amyotrophic lateral sclerosis	432
16	Crohn's disease	385
17	inflammatory bowel disease	355
18	long QT syndrome	345
19	chronic obstructive pulmonary disease	332
20	ulcerative colitis	330

Table 2.6: HDO annotation produced by *OntoSuite-Miner* from the Ensembl variation database. The top 20 most annotated genes in the table on the left and the top 20 most annotated diseases in the table on the right. The 'Count' is the number of unique gene/disease a disease/gene is annotated with. Diseases like 'obesity', 'schizophrenia', 'bipolar disorder' and cancers were popular topics in GWAS studies, thus have most support in the Ensembl variation database. Interestingly, many uncharacterized gene(symbol starts with 'LOCL' followed by a number) were found and annotated with a lot of diseases. This is because, compared to the traditional candidate gene approach, GWAS studies usually identify a large number of SNPs for the disease under study, thus generating interesting and novel gene candidates that may have never been linked to the disease before.

Human Disease Gene Database (*HDGDB*)

The results from the above three data sources were merged into a comprehensive human disease gene data base, named *HDGDB*. The overlap between genes, diseases and gene-disease associations between the three sources are shown in fig. 2.10. I will introduce *HDGDB* in the following paragraphs from three perspectives, it's gene features, disease features and the gene-disease association(GDA) features. The validation of *HDGDB* are then performed both quantitatively and qualitatively and possible improvements are discussed.

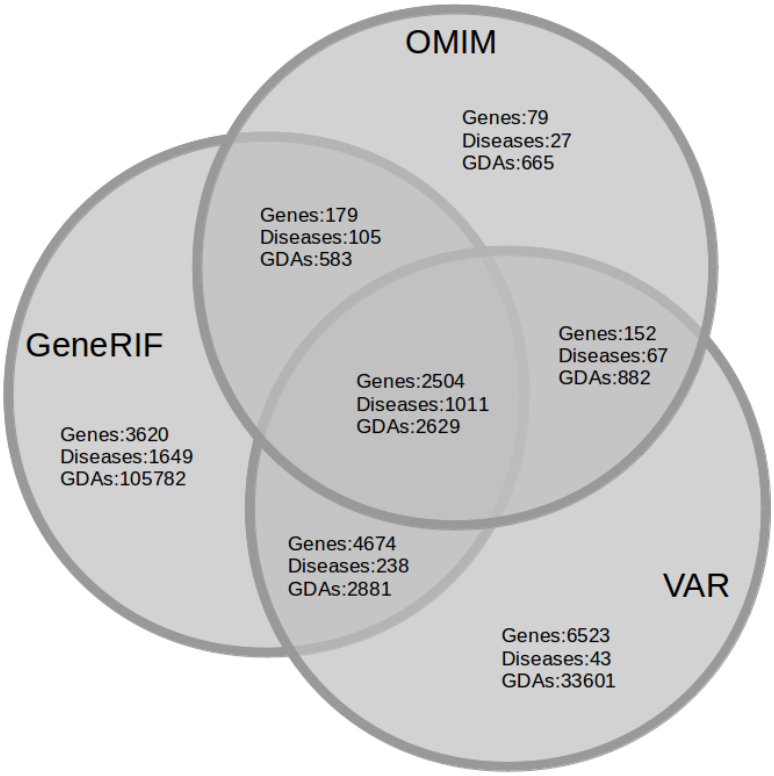


Figure 2.10: Venn diagram showing the overlaps between genes, diseases and GDAs in *HDGDB* according to source. Only 1.8% (2881) the GDAs are common to all of the data sources, while in the case of genes and diseases the overlap is 14% (2504) and 32.2% (1101) respectively. Such small overlaps between data sources highlights the importance of data integration.

Gene features

Reactome pathway database [172] provides pathway annotation for genes while the

Panther Protein Class Ontology classify genes (gene products) into different classes [173, 174]. Genes in *HDGDB* are thus classified with 1) 24 top-level pathways from Reactome database and 2) 29 top-level classes from Panther Protein Class Ontology to explore the composition of genes in terms of pathways and protein classes. The result is shown in fig. 2.11. The best-represented pathways are 'Signal Transduction' and 'Metabolism', comprising 10% of the disease genes each, followed by 'Immune System' (7%), 'Gene Expression' (5.5%) and 'Metabolism of proteins' (5%). In *HDGDB*, over 80% of genes (14306 genes) are protein-coding genes (calculated based on UniProt Swiss-Prot manually reviewed records on 1 Feb 2016). The remaining 20% belong to pseudogenes, ncRNA and other categories. There are estimated 19000 protein-coding genes in the human genome [164], of which roughly 75% are annotated with at least one disease in *HDGDB*. In terms of protein class, the best-represented are 'nucleic acid binding', comprising nearly 12% of all disease proteins followed by 'receptor' (9%), 'hydrolase' (8.5%), and 'transcription factor' (8.2%). The next best-represented protein classes are 'enzyme modulator', 'transferase' and 'signaling molecule' comprising approximately 7% each. Note that more than half of the genes in *HDGDB* do not appear in any pathways and about 40% of the proteins encoded by the disease genes are not covered by the Panther Protein Class Ontology ('Unclassified'). The numbers are even higher in the Ensembl variation data. This is possibly due to the fact that the pathways and protein classes are manually annotated and have limited gene coverage. Some of the recent discoveries are not included in these manually curated datasets but captured by *OntoSuite-Miner*.

As shown in fig. 2.12a, most of the genes have only a few disease annotations while only a few genes have been linked to a large number of diseases. The same can be observed when analyzing diseases and their associated genes (fig. 2.13a). The most well-annotated genes in *HDGDB* are 'TNF tumor necrosis factor' (Entrez id 7124), 'VEGFA vascular endothelial growth factor A' (Entrez id 7422) and 'TP53 tumor protein p53' (Entrez id 7157) which have been linked to 427, 414 and 401 diseases respectively. The 'TNF' gene is involved in the regulation of a wide range of biological processes including cell proliferation, differentiation, apoptosis, lipid metabolism, and coagulation and implicated in a variety of diseases, including autoimmune diseases, insulin resistance, and cancer [175, 176]. The 'VEGFA' gene is up-regulated in many known tumors and its expression is correlated with tumor stage and progression [177,178]. The 'TP53' gene is a well-studied tumor suppressor where it's encoded proteins bind to DNA and regulate other gene expressions to prevent mutations. It has

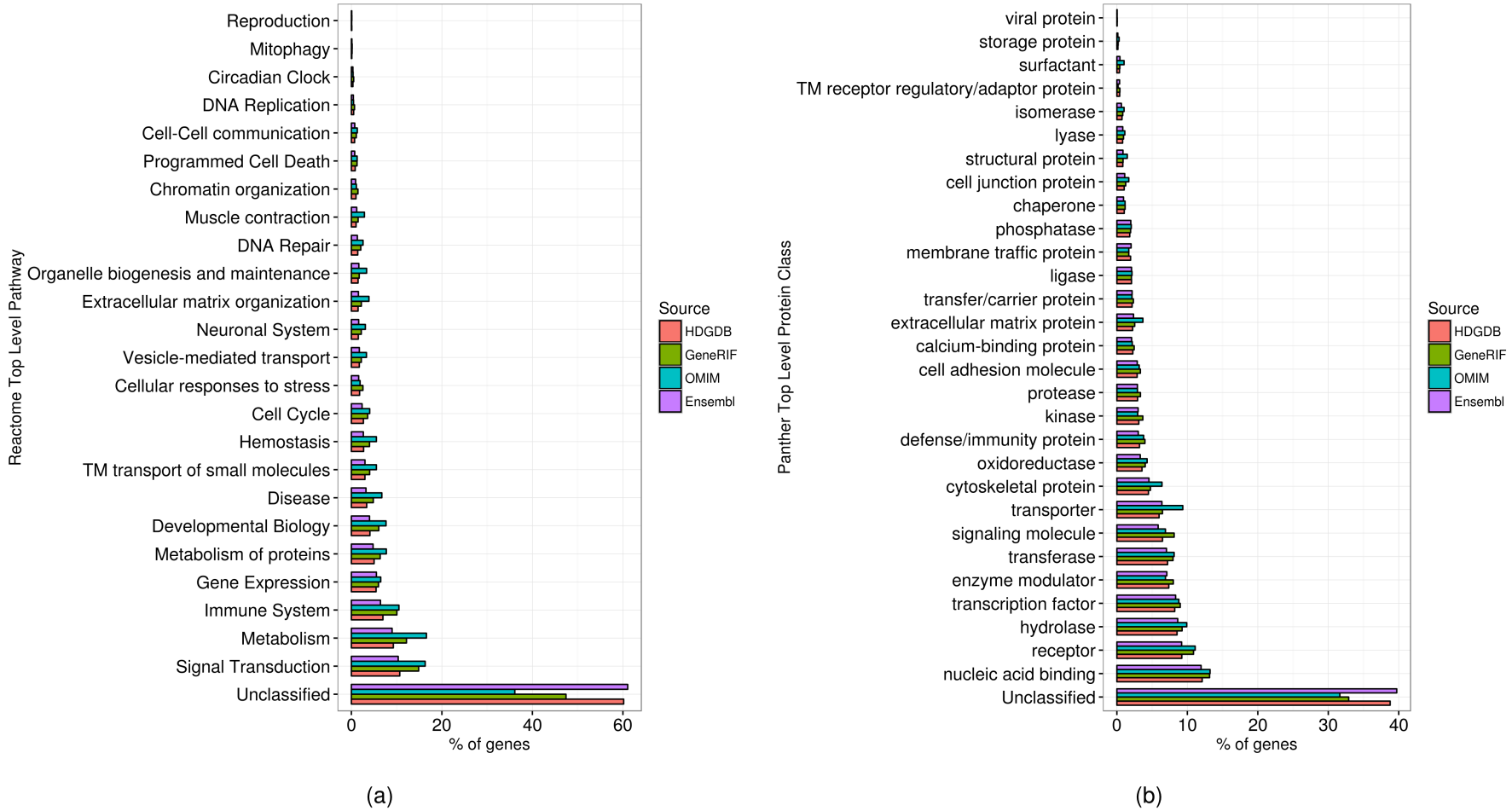
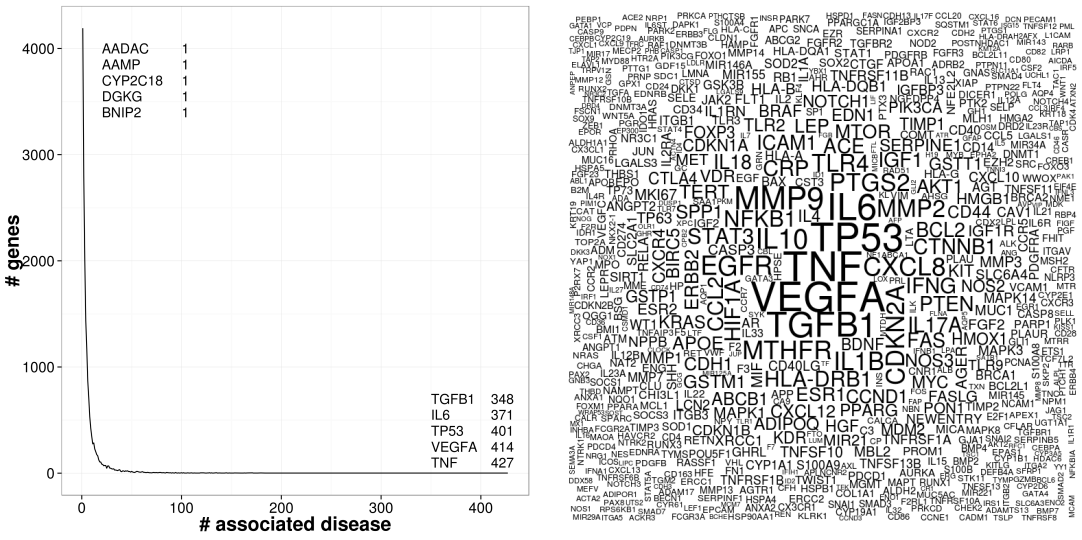


Figure 2.11: Distribution of genes in $HDGDB$ by (a) top level Reactome pathways and by (b) top level Panther protein classes. 'Signal Transduction' is the best-represented pathways while 'nucleic acid binding' is the best-represented protein class in $HDGDB$ genes. More than half of the genes do not appear in any pathways and about 40% of the proteins encoded by the disease genes are not covered by the Panther Protein Class Ontology('Unclassified'). This is partly because $HDGDB$ contains disease annotation for roughly 75% of the estimated 19000 protein-coding genes in the human genome [164], but the protein class annotation and pathway annotation only cover 11139 (58.6%) and 8613 (45.3%) human genes. The manual annotation process limited their gene coverage, thus result in a large number of genes being 'Unclassified'.

been classified as the most frequently mutated gene (>50%) in human cancer, indicating its crucial role in cancer formation [179]. In total, on average 8 diseases were annotated to each gene in HDGDB.

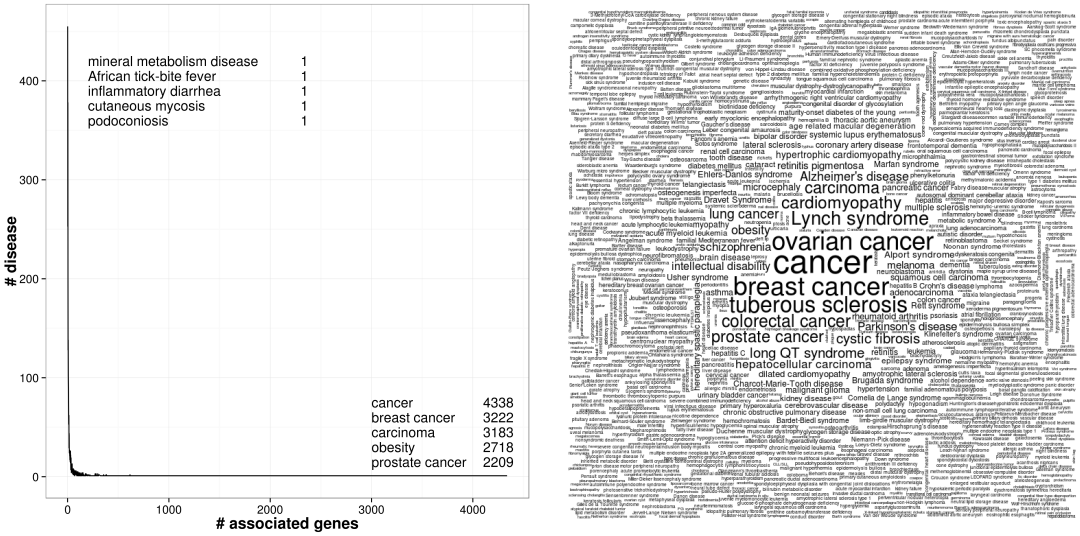


(a) (b)

Figure 2.12: (a) Distribution of the number of associated HDO terms per gene in *HDGDB*. Most of the genes have only a few disease annotations while a few genes have been linked to a large number of diseases. (b) Word cloud plot of the number of disease annotation for each gene in *HDGDB*.

Disease features

Genes are annotated with disease terms from HDO. Similar to the above gene analysis, most of the diseases were annotated with only a few genes while a few diseases (mostly cancers) were linked to a large number of genes (fig. 2.13a). The distribution of top-level HDO terms in the *HDGDB* is shown in fig. 2.14. The largest disease category is 'disease of anatomical entity'. 70% of the genes are annotated within this category. The second largest disease categories are 'disease of cellular proliferation', covering 57% of the genes, 'disease of metabolism' and 'disease of mental health' annotated to 27% of the genes each. 'syndrome' and 'physical disorder' are the least annotated disease category being only 5% each. In terms of individual disease, the top annotated diseases are mostly cancers, for example, 'breast cancer' has been linked to over 3000 genes and 'prostate cancer' has been linked to 2000. Other diseases like 'obesity', 'schizophrenia' and 'Alzheimer's disease' are also near the top of the list.



(a) (b)

Figure 2.13: Distribution of the number of genes associated with each HDO term in *HDGDB*. Most of the diseases have only a few gene annotations while a few diseases have been linked to a large number of genes. (b) Word cloud plot of the number of annotated genes for each HDO term in *HDGDB*

Gene-Disease association feature

The current release of *HDGDB* contains 147023 unique GDAs identified by *OntoSuite-Miner* from three sources including GeneRIF, OMIM and Ensembl variation. GDAs supported by multiple sources and/or identified by both annotators are more reliable than those with less supporting evidence. In order to indicate the level of confidence in each association, a confidence score is implemented. The score takes into account the number of data sources (GeneRIF, OMIM and/or Ensembl variation) that report the association, the number of evidences (text corpora reporting the GDA) in each sources and the number of annotators reporting the annotation. It is calculated as follows:

Let S be the score for a GDA.

$$S = \sum S_{source} = S_{OMIM} + S_{GeneRIF} + S_{Ensemblvariation} \tag{2.3}$$

The score for an individual source is defined:

$$S_{source} = W_{source} \cdot f(n) \tag{2.4}$$

where n is the number of evidence (text corpora reporting the GDA) in the source. The

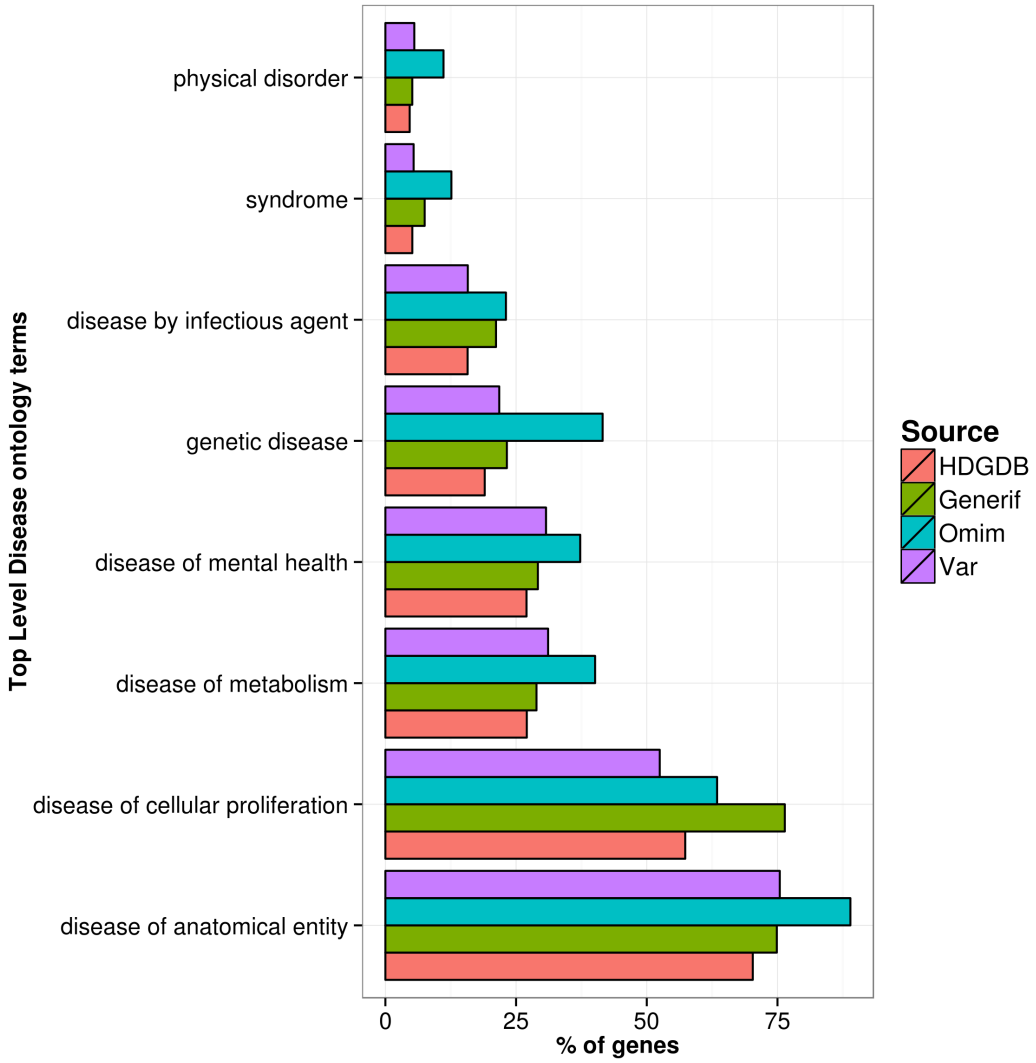


Figure 2.14: Disease category distribution of all diseases in $HDGDB$ using the top-level HDO terms. 'disease of anatomical entity' is the largest disease category, annotated with 70% of all the genes in $HDGDB$ while 'syndrome' and 'physical disorder' are the least annotated disease categories (5% each).

sources are equally weighted in the current implementation:

$$W_{source} = W_{OMIM} = W_{GeneRIF} = W_{Ensemblvariation} = \begin{cases} 0.2 & \text{if GDA is from } MeM \text{ and } NcA \\ 0.06 & \text{if GDA is from } MeM \text{ or } NcA \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

The weight of the annotator was designed so that a) W_{source} is always smaller than 1 and b) annotations found by two annotators were scored higher than those found by one annotator. The function $f(n)$ is an arbitrary increasing function depend on the number of evidence n in the source. It is designed to distinguish the existence of evidence while preventing the score being dominant by the unbalanced number of evidence across different sources. The rationale behind $f(n)$ is that a GDA is more reliable if it is supported by multiple sources than a single source for multiple times. Thus, $f(n)$ increases very fast when n is small while creating a long tail toward 1. k can be adapted to control the influence of n to a desired degree. For example, as shown in fig. 2.15, when $k = 0.2$, $f(n)$ start from 0.83, which means when there is any evidence exists, 83% of the corresponding source weight is counted. As n increase, this percentage gradually approach to 1. In the current implementation, k is set to be 0.2.

$$f(n) = \frac{n}{n+k} \tag{2.6}$$

Given the above equations, for example, gene *A2M* was found to be linked to 'DOID:10652 Alzheimer's disease' in HDGDB. Multiple sources of evidence exist and the number of text corpora reporting this GDA in different sources is shown in the following matrix:

$$A = \begin{array}{c} \\ OMIM \\ GeneRIF \\ VAR \end{array} \begin{array}{c} \begin{array}{ccc} M & N & M\&N \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 5 & 1 & 7 \\ 0 & 0 & 2 \end{pmatrix} \end{array}$$

The score for this GDA is calculated as follow:

$$S_{OMIM} = 0.2f(1) = 0.167$$

$$S_{GeneRIF} = 0.2f(7) + 0.06f(1) + 0.06f(5) = 0.302$$

$$S_{VAR} = 0.2f(2) = 0.182$$

$$S = \sum S_{source} = S_{OMIM} + S_{GeneRIF} + S_{VAR} = 0.651$$

The GDA score ranges from 0 to 1. It can be used to assist in the prioritization, weighting and navigation of the GDAs since it indicates their level of confidence. For

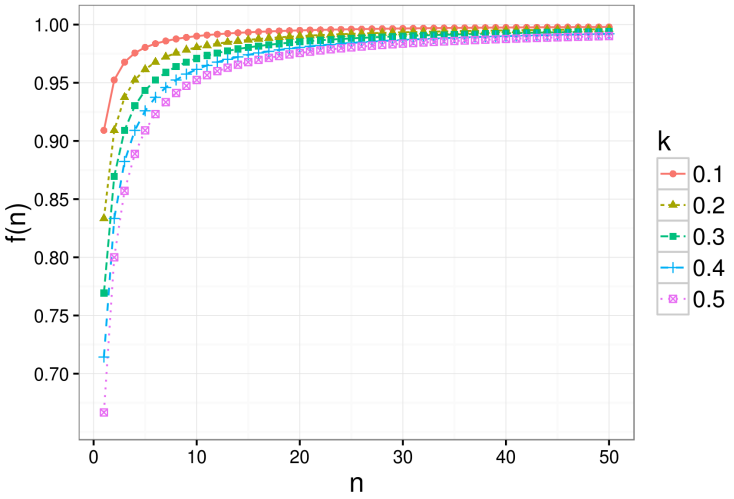


Figure 2.15: How k affect the source score $f(n) = \frac{n}{n+k}$. The function $f(n)$ is an arbitrary increasing function depend on the number of evidence n in the source. It is designed to distinguishe the existence of evidence while preventing the score being dominant by the unbalanced number of evidence across different sources. $f(n)$ increases very fast when n is small while creating a long tail toward 1. k can be adapted to control the influence of n to a desired degree. In the currently implementation, k is set to be 0.2.

example, associations found in one source by both annotators ($0.166 \leq S \leq 0.2$) have higher scores than those only found by one annotator($0.083 \leq S \leq 0.1$). Associations found by both annotators in two sources($0.332 \leq S \leq 0.4$) have a higher score than those found in a single source.

The sources were weighted equally in the currently implementation. The gene-disease associations identified in human curated data sources such as OMIM, are likely to be more precise than sources such as GeneRIF. However, the latter may provide useful information for genes when there is no curated data available or for those that are very recently identified. Thus, the annotation score could be refined base on the confidence/quality of the source, but it is unclear how to assign the weights to the sources when there is no obvious training data available.

In *HDGDB*, most of the GDAs were scored at around 0.2 which indicated that most of them were found only in a single source. This is because the focus of the three data sources are different. OMIM focuses on human genes and genetic phenotype; Ensembl variation database contains mainly GWAS studies for cancers while GeneRIF is a bit of both but focuses more on gene function. Their differing objective for describing gene-disease relationships may result in a low overlap between sources but

does not imply error. For example, *NAT*2 is linked to 'DOID:9352 type 2 diabetes mellitus' with a score of 0.05, having only one supporting evidence from a relatively recent paper published in 2013 [180] and contributing to *HDGDB* by a from GeneRIF entity. This association is not captured in OMIM nor Ensembl variation but it is a correct association. In fact, it is these low sored GDAs that are more likely to bring new insight into the understanding of complex diseases. On the other hand, 4% (5979) of GDAs scored over 0.3, supported by at least two sources. The top-50 scoring GDAs (table 2.7) are very well-studied gene-disease associations. For example, 'retinitis pigmentosa' with RPGR (retinitis pigmentosa GTPase regulator); 'Alzheimer's disease' with *PSEN*1 (Presenilin-1), *APP* (Amyloid precursor protein) and *APOE* (Apolipoprotein E); 'Wilson disease' with *ATP*7*B*. The top ranked GDA in *HDGDB* is the association of 'breast cancer' and *BRCA*2 (breast cancer 2) gene. *BRCA*2 is a tumor suppressor gene found in all humans. It encodes a protein, called by the synonym breast cancer type 2 susceptibility protein, is responsible for repairing DNA or destroying cells if DNA cannot be repaired. If *BRCA*2 itself is damaged by a mutation, damaged DNA is not repaired properly, and this increases the risk for breast cancer [181]. The second ranked GDA is the association of 'pulmonary hypertension' and *BMPR*2 (Bone morphogenetic protein receptor type II) gene. It has been shown that *BMPR*2 mutations are present in more than 70% of familial cases of relatives of patients with idiopathic pulmonary hypertension [182]. *BMPR*2 functions to inhibit the proliferation of vascular smooth muscle tissue. When it is inhibited, vascular smooth muscle proliferates and can result in 'pulmonary hypertension'. The association of 'melanoma' and *CDKN*2*A* (cyclin-dependent kinase Inhibitor 2A) gene ranked the third in HDGDB. *CDKN*2*A* codes two proteins, $p16$ and $p14arf$, which act as tumor suppressors. $p16$ binds to the cyclic dependent kinases CDK4 to inhibit their ability to create tumors, but when inactivated the suppression no longer occurs [183], thus starting the development of melanoma. *CDKN*2*A* is estimated to be the second most commonly inactivated gene in cancer after $p53$ and it's mutation has been associated with melanoma in many previous studies [184–186]. These GDAs were supported by multiple evidences from different sources and therefore obtained the highest scores.

The confidence score is especially useful in annotation generated by automatic methods such as the *HDGDB* by *OntoSuite-Miner*. It provides a reference to rank/weight the associations based on evidence and assists in the prioritization and navigation of the GDAs. In addition, the score can be applied in analysis such as gene set enrichment analysis (discussed in the next chapter) so that those highly ranked GDAs are to

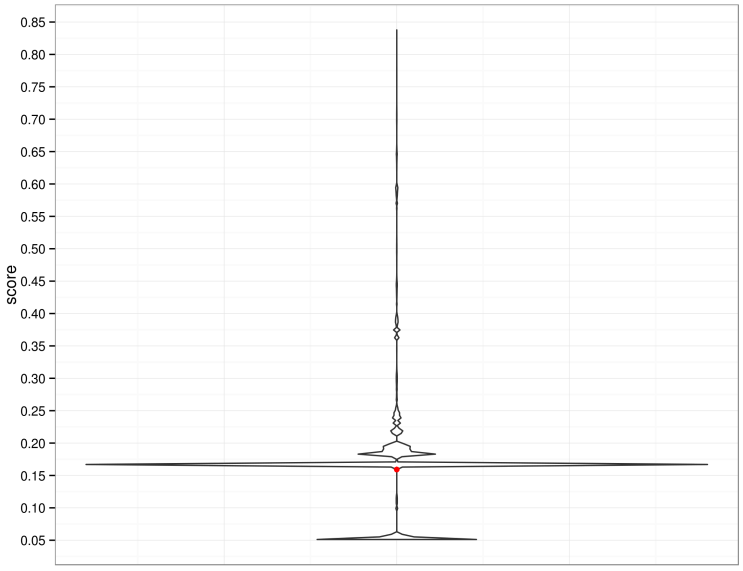


Figure 2.16: Violin plot of the distribution of GDA scores in *HDGDB*. 64% of the GDAs were scored at within the range of [0.15,0.2] which indicated that they were found only in a single source, indicating the small overlap between the three sources, namely OMIM, GeneRIF and Ensembl variation.

be weighted more then other.

OMIM did not provide data linking their entities to literature but all the entities in GeneRIF and most of the entities in Ensembl variation database have references (usually papers in PubMed database) attached as supporting evidence. Thus, Almost all of the GDAs (98.47%) in *HDGDB* can be linked back to the sources and the corresponding papers. For those GDAs that have no references, the original entities in the corresponding sources are provided. Remarkably, 82.48% of the articles supporting the GDAs in *HDGDB* have been published in the last 10 years, 58.74% in the last 5 years (fig. 2.17). This indicates that *OntoSuite-Miner* is able to pick up GDAs from some of the most recent studies.

2.3.3 Updating HDGDB

HDGDB is a gene disease association database containing GDAs from OMIM, GeneRIF and Ensembl variation databases. table 2.8 displays the statistics of the current database, v3.0, generated on 2016.01.05, compared with the previous ones v1.0 on 2014.09.04 and v2.0 on 2015.02.17. As shown in fig. 2.18, the OMIM database only has a small change because of the time consuming nature of the expert-curation processes.

	DOID	Disease	Gene	Score	Score.o	Score.g	Score.v	#Evidence(o/g/v)
1	DOID:1612	breast cancer	BRCA2	0.84	0.217	0.319	0.259	2/273/2466
2	DOID:6432	pulmonary hypertension	BMPR2	0.83	0.217	0.312	0.257	2/48/34
3	DOID:1909	melanoma	CDKN2A	0.83	0.232	0.313	0.258	3/87/43
4	DOID:1099	alpha thalassemia	HBA1	0.83	0.217	0.31	0.256	2/26/32
5	DOID:10584	retinitis pigmentosa	RPGR	0.83	0.217	0.309	0.253	2/26/27
6	DOID:2352	hemochromatosis	HFE	0.78	0.167	0.318	0.256	1/105/29
7	DOID:893	Wilson disease	ATP7B	0.78	0.167	0.307	0.259	1/63/141
8	DOID:3490	Noonan syndrome	PTPN11	0.78	0.167	0.307	0.259	1/40/114
9	DOID:1485	cystic fibrosis	CFTR	0.78	0.217	0.316	0.2	2/289/1186
10	DOID:0050773	paraganglioma	SDHD	0.78	0.217	0.254	0.256	2/25/51
11	DOID:0050773	paraganglioma	SDHB	0.78	0.217	0.256	0.253	2/34/21
12	DOID:1099	alpha thalassemia	HBA2	0.77	0.217	0.31	0.256	2/21/32
13	DOID:9253	gastrointestinal stromal tumor	KIT	0.77	0.167	0.315	0.253	1/93/14
14	DOID:9255	frontotemporal dementia	VCP	0.77	0.217	0.254	0.255	2/16/18
15	DOID:9253	gastrointestinal stromal tumor	PDGFRA	0.77	0.167	0.313	0.255	1/46/19
16	DOID:649	prion disease	PRNP	0.77	0.167	0.314	0.245	1/41/9
17	DOID:9255	frontotemporal dementia	TARDBP	0.77	0.167	0.312	0.249	1/44/16
18	DOID:14686	Axenfeld-Rieger syndrome	FOXC1	0.77	0.217	0.248	0.253	2/8/14
19	DOID:13628	glucose-6-phosphate dehydrogenase deficiency	G6PD	0.77	0.188	0.307	0.257	3/36/46
20	DOID:3012	Li-Fraumeni syndrome	TP53	0.77	0.167	0.305	0.258	1/21/92
21	DOID:4252	Alexander disease	GFAP	0.77	0.167	0.303	0.257	1/36/32
22	DOID:1909	melanoma	MITF	0.77	0.167	0.31	0.217	1/50/2
23	DOID:2739	Gilbert syndrome	UGT1A1	0.77	0.167	0.306	0.253	1/15/14
24	DOID:1921	Klinefelter's syndrome	FGFR1	0.76	0.167	0.292	0.257	1/7/39
25	DOID:0050773	paraganglioma	SDHC	0.76	0.217	0.248	0.24	2/8/7
26	DOID:10632	Wolfram syndrome	WFS1	0.76	0.217	0.248	0.255	2/22/19
27	DOID:14261	fragile X syndrome	FMR1	0.76	0.217	0.305	0.182	2/53/2
28	DOID:0050589	inflammatory bowel disease	NOD2	0.76	0.167	0.304	0.251	1/17/12
29	DOID:3612	retinitis	RPGR	0.76	0.217	0.247	0.241	2/24/27
30	DOID:0060241	3-M syndrome	CUL7	0.76	0.167	0.282	0.254	1/4/15
31	DOID:0060472	Kindler syndrome	FERMT1	0.76	0.167	0.296	0.244	1/13/6
32	DOID:13636	Fanconi's anemia	FANCA	0.76	0.167	0.296	0.248	1/11/20
33	DOID:11105	fundus albipunctatus	RLBP1	0.75	0.217	0.236	0.239	2/4/6
34	DOID:9993	hypoglycemia	ABCC8	0.75	0.217	0.225	0.242	2/7/51
35	DOID:0050439	Usher syndrome	PDZD7	0.74	0.217	0.236	0.217	2/4/2
36	DOID:10652	Alzheimer's disease	APOE	0.72	0.167	0.319	0.198	1/301/20
37	DOID:10652	Alzheimer's disease	APP	0.72	0.167	0.32	0.199	1/420/28
38	DOID:1612	breast cancer	ESR1	0.72	0.167	0.318	0.196	1/373/11
39	DOID:10652	Alzheimer's disease	PSEN1	0.72	0.188	0.318	0.199	3/133/69
40	DOID:1926	Gaucher's disease	GBA	0.72	0.192	0.257	0.259	5/42/161
41	DOID:1307	dementia	MAPT	0.72	0.167	0.258	0.256	1/55/27
42	DOID:1612	breast cancer	TP53	0.72	0.167	0.317	0.167	1/221/1
43	DOID:1919	Lesch-Nyhan syndrome	HPRT1	0.72	0.167	0.249	0.258	1/9/57
44	DOID:8997	polycythemia vera	JAK2	0.72	0.167	0.317	0.182	1/86/2
45	DOID:10487	Hirschsprung's disease	RET	0.72	0.167	0.258	0.258	1/48/68
46	DOID:3490	Noonan syndrome	SOS1	0.72	0.167	0.251	0.258	1/12/59
47	DOID:10283	prostate cancer	AR	0.71	0.167	0.315	0.196	1/280/10
48	DOID:9256	colorectal cancer	TP53	0.71	0.167	0.316	0.188	1/81/3
49	DOID:9263	homocystinuria	CBS	0.71	0.167	0.251	0.257	1/12/52
50	DOID:14330	Parkinson's disease	SNCA	0.71	0.182	0.314	0.198	2/236/23

Table 2.7: The top 50 scored gene disease associations in the Human Disease Gene Database

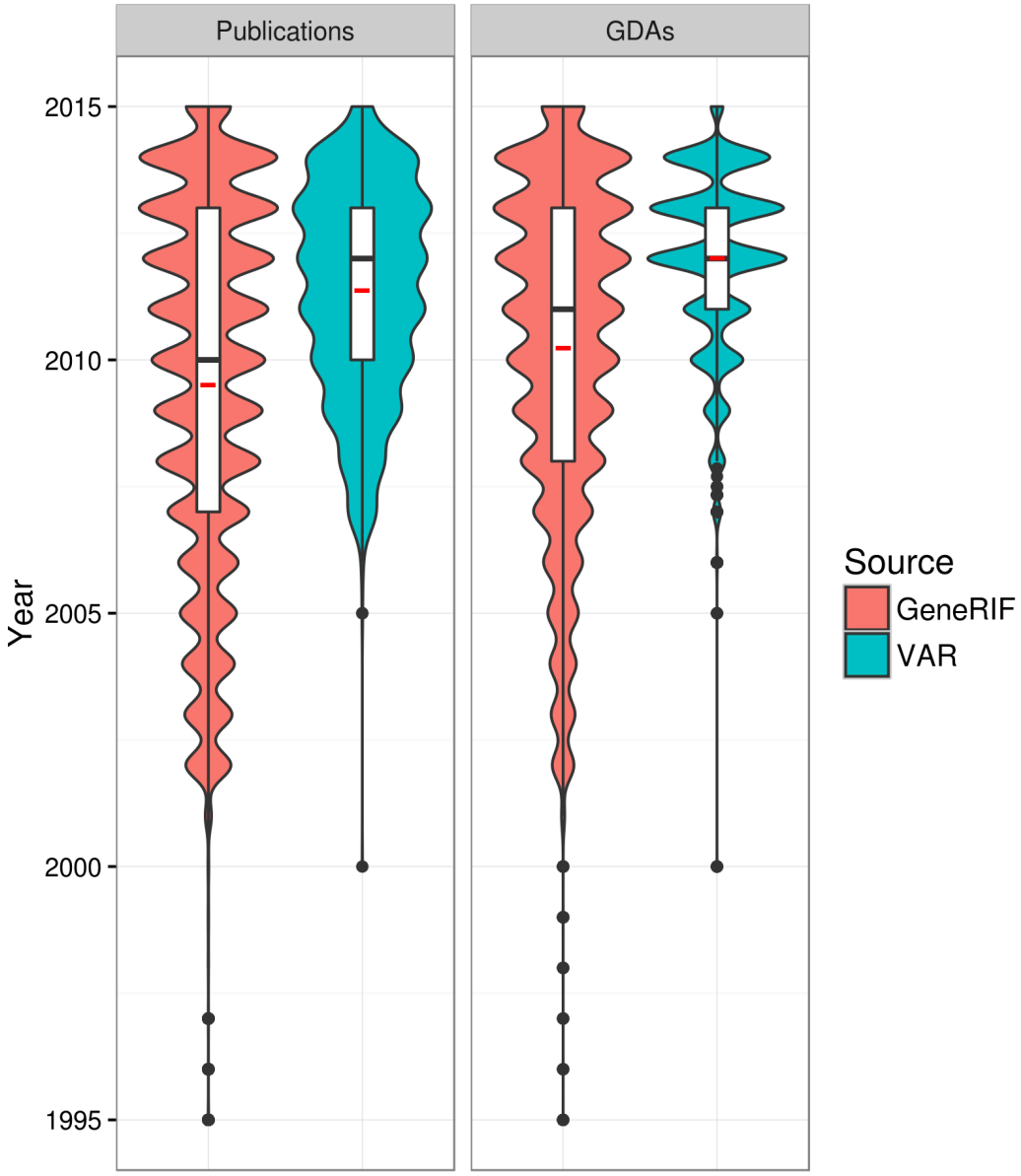


Figure 2.17: Violin plot of the distribution of publication year of the publications in *HDGDB* and their supported GDAs in GeneRIF and VAR. Any papers published before 1995 were binned as 1995. OMIM does not provide a reference for their data, thus it was not included in the figure. The red line in the box plot indicates the mean. The paper time line clearly shows that the GeneRIF dataset has been continually increasing since 2001 while the Ensembl variation database started to expand from 2007 when Genome Wide Association studies became more popular. In total, 82.48% of the GDAs were supported by papers from the last 10 years.

GeneRIF has increased constantly. A sharp increase of the Ensembl variation database between v1.0 to v2.0 is the result of an major update which added in a number of recent GWAS studies. Despite the number of entities in each source increasing, a small drop in GDAs were observed. This is because of the implementation of a filter (see section 2.2.4) between v2.0 and v3.0 which removed some of the erroneous GDAs to increase mapping accuracy. For example, in v2.0, a GeneRIF *Overexpression of serum A1BG is associated with non-small cell lung cancer* was mapped to both 'lung cancer' and 'cancer' creating two GDAs due to an annotation error referred to as a coordinating conjunctions(see section 2.3.4.1). The filter removed the latter, keeping only the most specific annotation 'lung cancer'. Thus the total GDAs decreased but the accuracy was improved. *OntoSuite-Miner* allows regular updating of the data sources, thus keeping the database up-to-date with the most recent findings.

source	Genes			Diseases			Associations			# corpora in source		
	v1.0	v2.0	v3.0	v1.0	v2.0	v3.0	v1.0	v2.0	v3.0	v1.0	v2.0	v3.0
Generif	10376	10466	10977	2774	2810	3003	99393	112818	111875	397403	439114	484340
OMIM	2628	2770	2914	1057	1142	1210	4244	5146	4759	5339	5504	5724
Var	4829	13980	13853	385	1063	1359	8452	39133	39993	34790	229777	337731
HDGDB	12964	17547	17731	2841	2917	3140	108010	147823	147023	563141	674395	827795

Table 2.8: The number of gene, disease and gene-disease association between different version of *HDGDB*.

2.3.4 Validation of HDGDB

In order to validate the *HDGDB*, I evaluated the gene disease association data in several ways. First, I manually reviewed 900 HDO mappings randomly taken from *HDGDB*, 300 from each source including OMIM,GeneRIF and Ensembl variation. Precision rates were calculated to measure the accuracy of the dataset. Errors/mismatches were identified and classified into four different groups and possible solutions were discussed. Next I compared *HDGDB* against an OMIM 'gold standard' dataset generated by using HDO term's cross references features (see section 2.3.4.2). Finally, two well studied gene sets, an aging related gene set from GenAge database [187] and a ciliopathy related gene set from Cildb [188], were examined with *HDGDB*, to explore their gene disease relations by running over-represented disease ontology enrichment analysis.

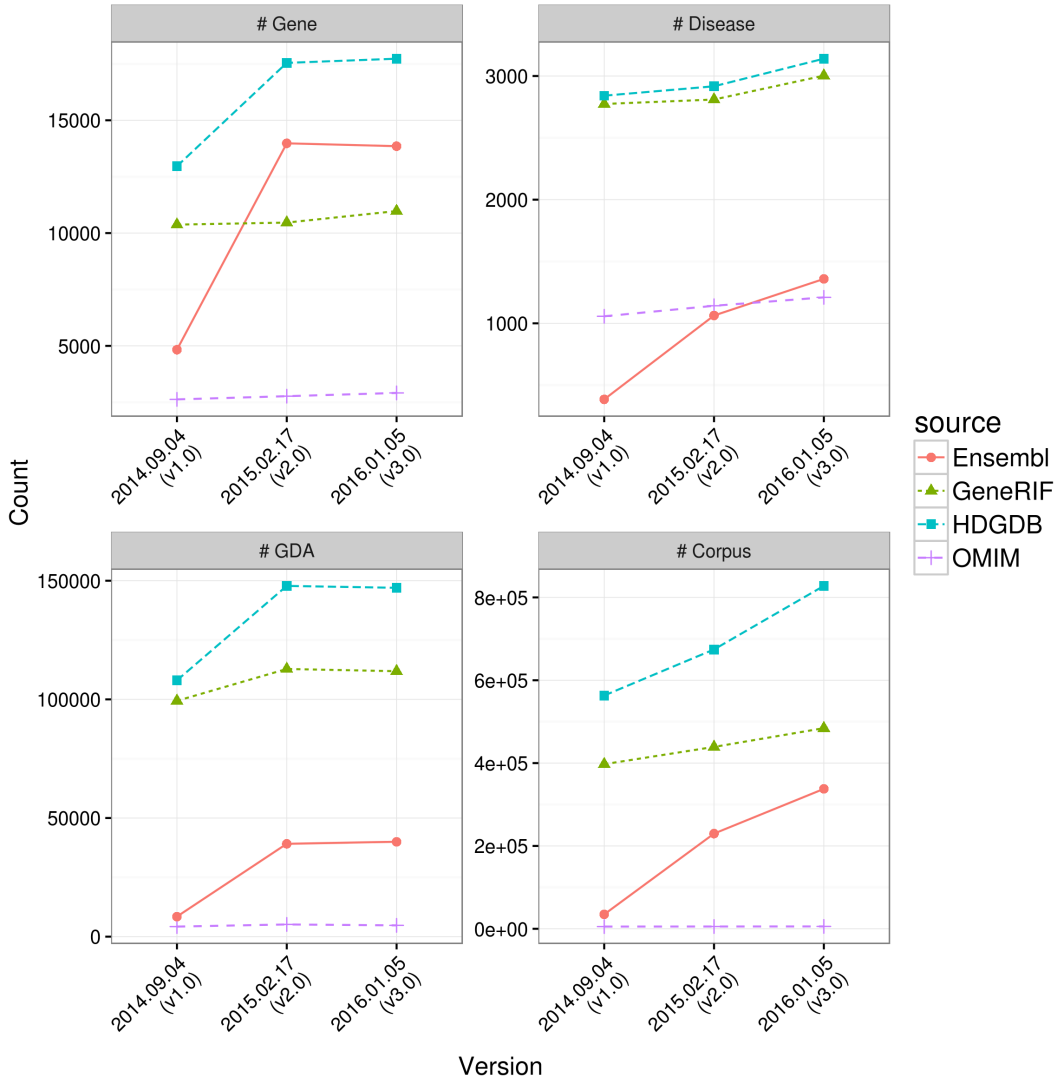


Figure 2.18: Number of gene, disease, GDA and the number of corpus between major updates. A small drop in GDAs were observed. This is because of the implementation of a filter (see section 2.2.4) between v2.0 and v3.0.

2.3.4.1 Dealing with annotation errors

Annotation errors from *OntoSuite-Miner* may arise from a variety of sources including the two NLP annotators used in the task, the ontology itself and errors induced by the original text sources. Some errors are inevitable in an automated process like *OntoSuite-Miner* but some of them can be avoided or corrected. Due to the lack of a standard labeled HDO annotation, there is not an obvious way to evaluate the quality of *HDGDB*. As a rule of thumb, manual inspection is the most accurate way to evaluate the performance of NLP based methods. Thus, I reviewed 900 HDO mappings (referred to as *ALL*) randomly taken from *HDGDB*, 300 from each source including OMIM, GeneRIF and Ensembl variation (not available in the thesis but provided as supplementary file on the attached disk). As shown in fig. 2.19, a small proportion of these 900 mapping were found solely by one of the annotators while the majority were found by both of the annotators. This applied to the results across all of the three data sources and *ALL*, suggesting that most of the time, the two annotators perform similarly.

By manually inspecting the accuracy of the mapping, precision rates were calculated for mappings from each of the data sources and for the overall 900 mappings as a whole. The precision rate is calculated as follows:

$$Precision = (Number\ of\ correct\ mapping)/(Total\ number\ of\ mappings) \quad (2.7)$$

It is difficult to estimate the recall rate in this case since it is extremely time consuming and requires the involvement of domain experts. Therefore, only precision rates are measured and used for the purposes of evaluation. As shown in fig. 2.20, the highest precision rates are from those mappings generated by both *MeM* and *NcA*, 95.15% in *ALL*. In terms of the performance of each annotator, *MeM* slightly out performed *NcA* across all sources. However, when looking at those mappings generated solely by one annotator, *MeM* largely out performed *NcA*, correctly identifying 85% of the mappings in *ALL* while *NcA* only identified 55.05% of them. The result suggests that in the current configuration, *MeM* performs better in this specific task than *NcA*, while the overlap mappings found by both of them are the most reliable.

Each annotator has its own pros and cons due to the different underlying algorithms and configurations. They favour certain types of text input and tend to make mistakes for others. A pattern of errors was observed, some of which have been already elucidated in [189]. Four types of errors were identified among the 900 mappings inspected. The details are shown in fig. 2.21 and table 2.9.

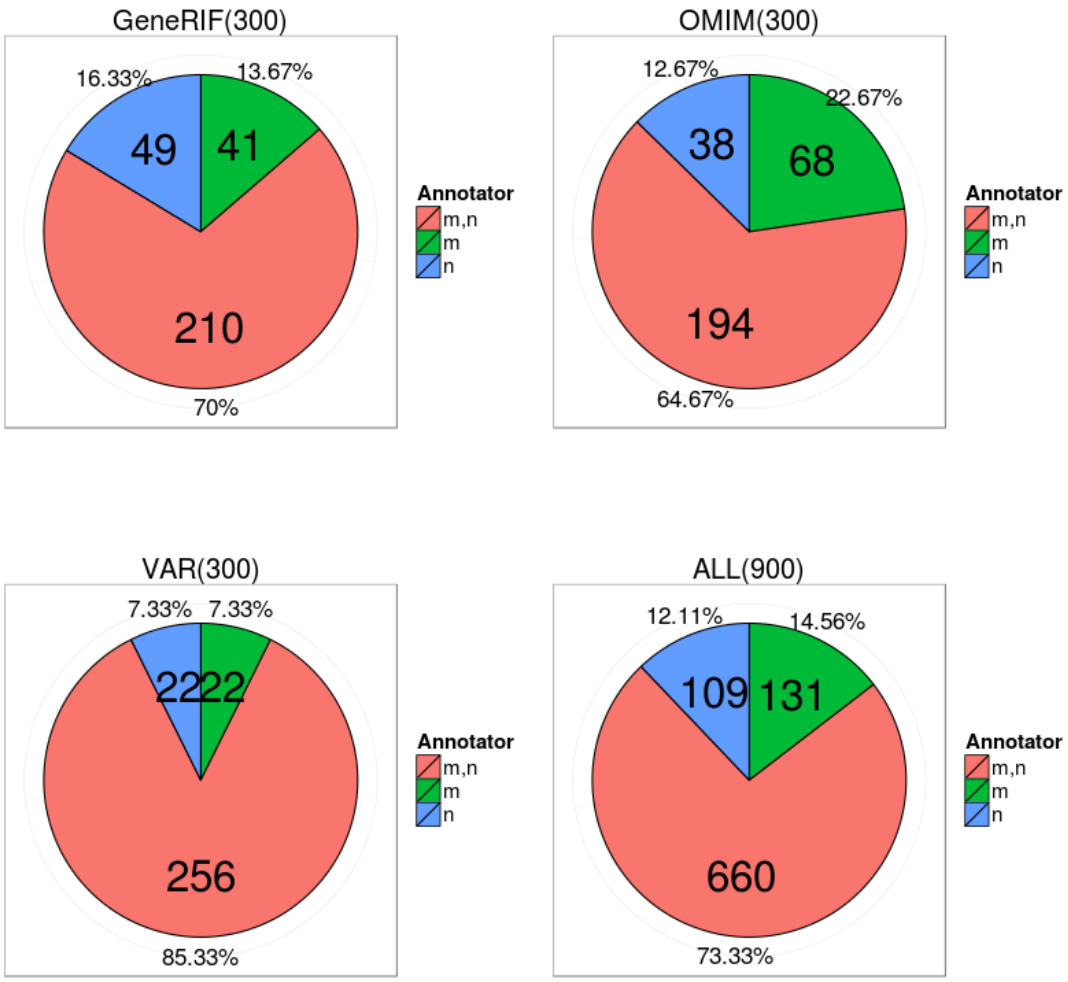


Figure 2.19: The distribution of annotations by annotator in the manually inspected 900 mappings from $HDGDB$. More than 70% of mappings were found by two annotators, indicating that most of the time, the two annotators' perform similarly and the generated mappings are trustworthy compared to those only found by one of them.

	MetaMap				NCBO Annotator				MetaMap &NCBO Annotator			
	GeneRIF	OMIM	VAR	ALL	GeneRIF	OMIM	VAR	ALL	GeneRIF	OMIM	VAR	ALL
CC	5	4	2	11	16	20	10	46	11	4	10	25
Abbr	2	1	1	4	4	0	0	2	2	0	0	4
Missing	0	0	0	0	0	0	0	0	0	1	0	1
Other	2	2	0	4	1	0	0	1	2	0	0	2

Table 2.9: he number of error identified in 900 mapping from the Human Disease Gene Database

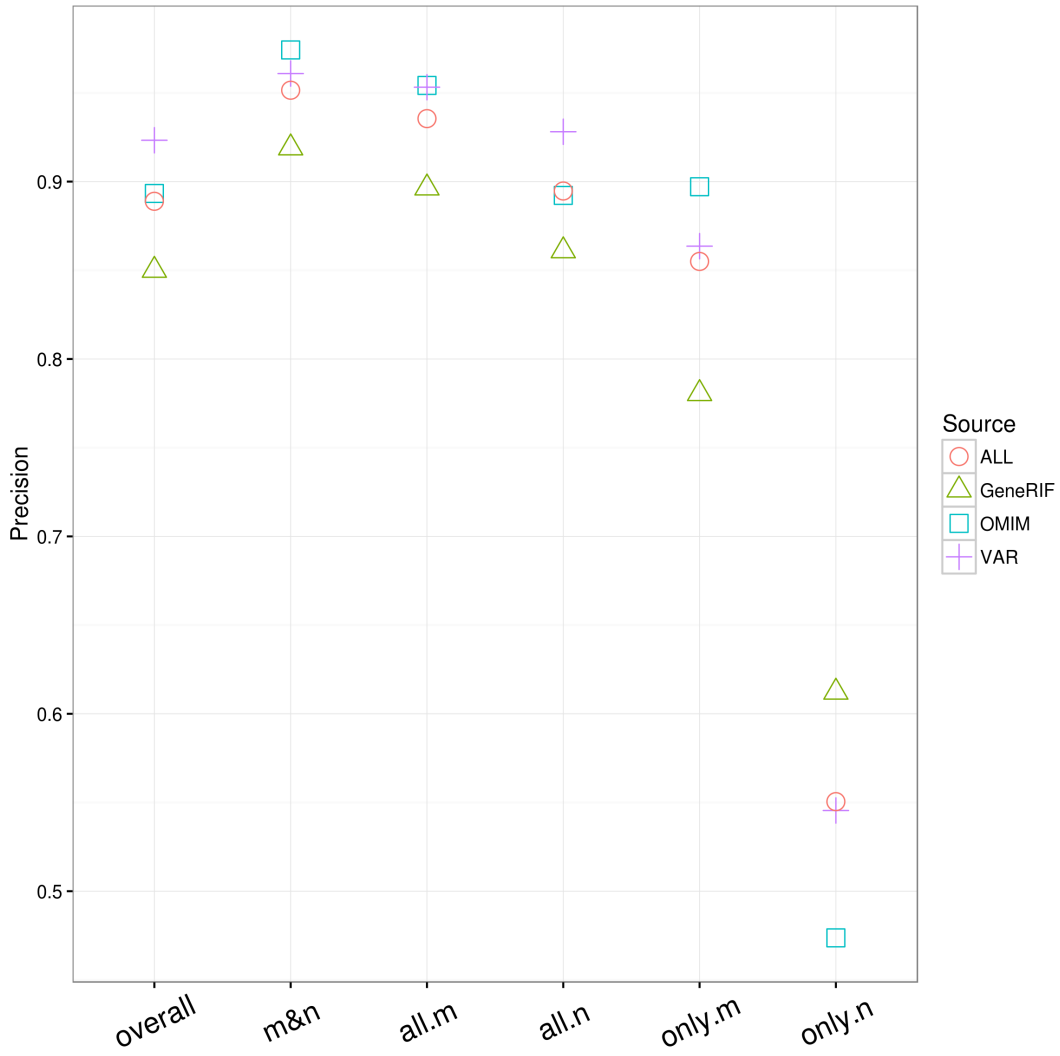


Figure 2.20: Precision rate of 900 manually inspected mappings from $HDGDB$. Precision rates were calculated based on the annotator that generated the mapping. The precision of those mappings generated solely by NcA or MeM were plotted as *only.n* and *only.m*. *all.n* and *all.m* indicates the precision of all mappings found by NcA or MeM respectively, including those found by both. Precision for the mappings found by both annotators is plotted as *m&n* while *overall* is the overall precision for the 900 mappings regardless of the annotator. Mapping source is indicated by shape while a *ALL* represent all 900 mappings. It is observed that mapping precision rates are the lowest when the mappings are only supported by NcA. In general, MeM out performed NcA in all three data sources. Those mappings supported by both MeM and NcA were found the most reliable. An overall 88.89% of mapping were found correct amount the 900 mappings.

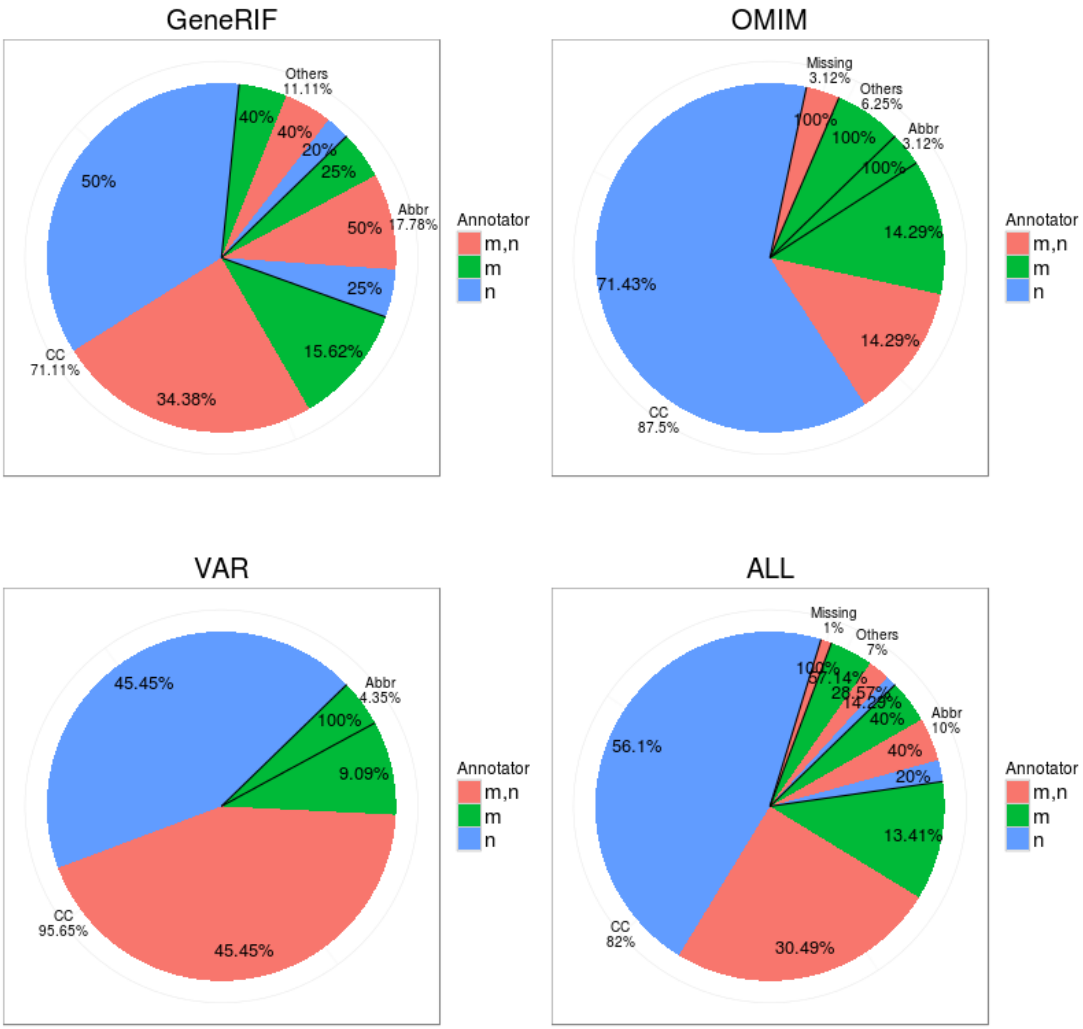


Figure 2.21: Distribution of the four types of annotation errors identified among the 900 mappings inspected from *HDGDB*. The most common type of errors, composed 82% of all the errors in *ALL*, were coordinating conjunctions (CC), the discovery of HDO terms with only a partial or ambiguous part of the input text. Other errors include those caused by abbreviation in the source text or errors due to missing concepts/synonyms in the Human Disease Ontology were identified but only composed 11% of all the errors in *ALL*. 7% of the errors were classified as 'Others', which were mainly caused by inappropriately handled prepositions of the source text or the inability to extract relationships between entities in the source text.

Mapping error due to Coordinating conjunctions The most significant problem contributing to the majority of errors (82% of all errors in the *ALL*) was the discovery of HDO terms which indicated a partial or ambiguous part of the input text. This type

of error, referred to as coordinating conjunctions in [189], occurs when the input text has complex syntax or contains an arbitrary way of stating multiple thoughts. This type of error can be further classified into two subcategories:

1)Wrongly split phrases. Some text have a complex syntax which confused the annotator's decision on how to split the text into phrases. This resulted in the using of the wrong parts of the text and consequently assigning the wrong HDO terms to the text. For example, *A protein encoded by this locus was found to be differentially expressed in postmortem brains from patients with atypical frontotemporal lobar degeneration* was wrongly assign to the HDO term 'DOID:1443 brain degeneration'. *Genetic variation may affect severity of disease for X-linked retinitis pigmentosa* was mapped to 'DOID:630 genetic disease' and *GLUTATHIONE SYNTHETASE DEFICIENCY OF ERYTHROCYTES HEMOLYTIC ANEMIA DUE TO* was mapped to 'DOID:13121 deficiency anemia'. Despite the fact that the correct HDO terms have been identified(not shown here) for the most of the text of this type, this was still a frequent source of mismatches.

2)Matching a partial of the sentence. This was the most prevalent source of errors, caused by recognizing only a partial of the sentence/phrase. Example including Dyserythropoietic anemia with thrombocytopenia being mapped to 'DOID:1588 thrombocytopenia' and *To analyze the expression of Syk tyrosine kinase during the multi step development of human breast carcinoma* being mapped to 'DOID:305 carcinoma'. This type of text shared a similar pattern. They either used a disease name as descriptive text(adjective), for example, *anemia with thrombocytopenia* to emphasize the condition/type of anemia, or include disease name that contain other disease names(usually more general disease), for example, *carcinoma* in *breast carcinoma*. Unfortunately, this pattern appear very frequently in medical text, thus making it one of the most prevalent source of errors in *HDGDB*. However, such error is not necessarily wrong in term of disease annotation. In the above case, it is reasonable to link the gene to 'thrombocytopenia' as well as 'anemia'. It is also true that the gene involved in the development of breast carcinoma is involved in carcinoma in general. Note that similar disease name does not always mean relevant disease. For example, *Charcot-Marie-Tooth disease, axonal, type 2V*, also known as 'CharcotMarieTooth(CMT) neuropathy' which is a nervous system disease characterized by progressive loss of muscle tissue and touch sensation across various parts of the body. The above text was wrongly assigned the term 'DOID:1091 tooth disease' which is a completely irreverent disease of dental disorder in mouth.

The HDO annotation was generated from sources including GeneRIF, OMIM and Ensembl variation, all of which provide short sentences/phrases to describe genes. Coordinating conjunctions will become more problematic when using longer, more complicated texts such as those in abstract/full-text documents of scientific literature. In terms of the performance of the two annotators, *MeM* tends to make the 'wrongly split phrase' mistake because it takes into account the gaps between words. In contrast, *NcA* only searches for exact matches, thus is unlikely to make such mistakes but is more likely to match a part of the sentence to HDO terms. In most cases in *HDGDB*, errors of this type result in less accurate matches that usually identify high level terms which are either direct or indirect ancestor terms of the correct term, for instance, 'DOID:162 cancer' and 'DOID:1612 breast cancer'. The effect of these errors varies depending on the type of analysis performed. For example, when constructing a gene-disease network, wrongly assigned high level terms may result in misleading high connectivity degree nodes. In other cases, such as testing for disease enrichment of a set of genes, these mappings are normally inconsequential to the overall enrichment result because high level terms will tend to have a large count in the background set and are therefore unlikely to be enriched in the study set (see chapter 3 on page 113 for detail).

An extra filtering process was implemented in the *OntoSuite-Miner* pipeline by taking into account the ontology hierarchy to minimize errors caused by coordinating conjunctions (see section 2.2.4 on page 53). Errors caused by coordinating conjunctions could also be rectified by analyzing the dependency grammar (see review [190]) of the text sentence. The idea is that the syntactic structure of a sentence consists of binary asymmetrical relations between the words of the sentence. A dependency parser can break the sentence into words connected according to their relationships. Such relationships can be visualized hierarchically where nodes represent words and edges represent relationships. As shown in fig. 2.22, *cat* is the syntactic subject of the sentence. *white* and *black eyes* are descriptive pieces of text (adjectival modifier or amod) that are used to modify/complete the meaning of *cat*. Such dependency trees can be used to guide the selection of the most relevant parts of the sentence. The words can be prioritized by their distance to the ROOT, that is the words that are closer to the ROOT have a higher priority. This word ranking can be fed into the annotators to avoid coordinating conjunctions to a certain degree.

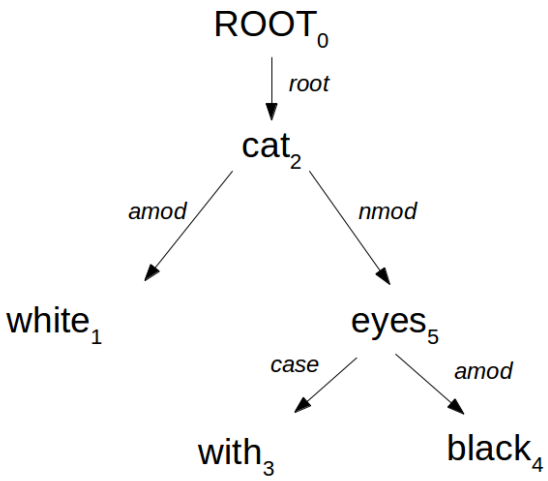


Figure 2.22: Dependency tree for the sentence *white cat with black eyes*. Each vertex in the tree represents a word (except ROOT), child nodes are words that are dependent on the parent. Edges are labelled by the relationship. The subscript indicates the position of the word in the sentence. The word *white* and *black eyes* are adjectival modifier (amod) and noun modifier (nmod) respectively which serve to modify the meaning of the noun phrase *cat*. In this example, *cat* is closer to the *ROOT*, indicating that it is the syntactic subject of the sentence. amod *white* is directly modifying the syntactic subject, thus has a closer relation to it than the nmod *eyes*. In this case, if to prioritized the words, *white* will have a higher priority than *eyes* and *cat with eyes*.

Mapping error due to Abbreviation The use of abbreviation in source texts results in annotators choosing the wrong HDO term because the meaning of the abbreviation is ambiguous. This can be further classified into three types of abbreviation error:

(1) gene/protein name wrongly interpreted as disease abbreviation. Examples including *WT1 expression is down-regulated in the ovulatory polycystic ovary syndrome endometrium* being wrongly mapped to 'DOID:5183 WT1 hereditary Wilmsś tumor' because of an exact synonym of the term being 'WT1', and, due to the same reason, *CRF does have a role to play in determining BDNF control of dendritic spines* being wrongly mapped to 'DOID:784 chronic kidney failure (CRF)' when it was referring to 'corticotrophin releasing factor' in the text.

(2) abbreviations being ambiguous when they are considered in a disease context. For example, *In amnion cells EGFR clustering induced by 50-Hz MF depends on acid sphingomyelinase activity* being mapped to 'DOID:8691 mycosis fungoides (MF)' when the *MF* means 'magnetic field'.

(3) ambiguous disease abbreviations. For example, *dlgap1 showed the highest association, of snps examined, with ocd* was wrongly annotated to 'DOID:84 osteochondritis dissecans (OCD)' when the *ocd* in the text was referring to 'obsessive-compulsive disorder'. *familial cold autoinflammatory syndrome 2* was incorrectly assigned the term 'DOID:3083 chronic obstructive lung disease (COLD)'. Another example would be *Emery-Dreifuss muscular dystrophy 2, AD* being mapped to 'DOID:10652 Alzheimer's disease(AD)' when the *AD* was short for 'autosomal dominant'.

These types of errors suggest that further research needs to be done into the development of the HDO. Great care needs to be taken when defining synonyms/concepts to avoid ambiguous/duplicated terms. In terms of approaches to improve the results, these types of errors can be reduced by applying a technique called WSD, that is word-sense disambiguation (see survey [191] for more details). WSD is a technique to identify the intended sense of a word (i.e. meaning) used in a sentence when the word has multiple meanings. It is still an open research area in NLP and different variants of WSD exist. The general procedure of WSD can be summarized as follows: given a set of words (e.g., a sentence or a bag of words), a technique is applied which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context. The knowledge source is a fundamental component of WSD, providing data which are essential to associate senses with words. For example, the word *AD* in *AD patients with an atypical onset have significantly higher levels of total tau* is more likely to refer to Alzheimer's disease due to the presence of the ward *tau*, which is a key protein involved in the disease. Thus, given enough knowledge sources, WSD might be able to distinguish ambiguous words' meaning in different contexts and assign the right HDO term to the text. *NcA* did not implement WSD but provides a parameter to specify the minimum match length which can solve some of the problems caused by abbreviations with the cost of possible loss of recall. However, it is unclear what the optimal match length is. A large minimum match length may result in a reduction of *NcA*'s detection power while a small one may not make any difference to the results. *MeM* has its own WSD procedure but does not provide any appropriate disease knowledge sources as training set for the WSD. The errors caused by abbreviations is likely to be reduced if when such training data become available.

Mapping error due to Missing concepts/synonyms in HDO There were also rare cases where a particular concept was not in the Human Disease Ontology (once in the 900 mappings examined). For example, *Auditory neuropathy, autosomal dominant, 1* representing Auditory neuropathy (AN), a variety of hearing loss in which the outer hair cells within the cochlea are present and functional, but sound information is not faithfully transmitted to the auditory nerve and brain properly. There is no HDO term for AN. The most relevant HDO terms were 'DOID:2742 auditory system disease' and 'DOID:870 neuropathy', of which the later was identified by both *MeM* and *NcA*. The lack of synonyms of HDO terms caused a lot of mismatches especially in disease contexts where disease names were highly variable in free text. For example, the word 'OCD' is frequently used as 'obsessive-compulsive disorder' but is not present as a synonym of that term. This type of error requires additional development research of the ontology to address. HDO is being actively updated, and new terms are being added to better represent human disease domain knowledge. The performance of the annotators is likely to be improved as this process continues. Data generated from *OntoSuite-Miner* identified such problem in ontologies, thus suggesting its potential role in aiding the developing/refining of ontologies.

Others There were several cases where an exact surface form in the text corpus was obvious but were missed by both *MeM* and *NcA*. For example, *ciliary dyskinesia primary 21* should be mapped to 'DOID:9562 primary ciliary dyskinesia' but neither of the annotators was able to find this term. This is likely caused by inappropriately handled prepositions. Another type of error occurs where some of the text describes a week, null or negative relationship between a gene and a disease, but the current implementation of *OntoSuite-Miner* does not capture this information, resulting in the generation of an incorrect gene-disease link. For example, *No association with ovarian cancer risk for BRCA1 or BRCA2 mutation carriers or with breast cancer risk for BRCA1 mutation carriers was observed* describes no association between the genes and ovarian cancer, but *OntoSuite-Miner* would generate a link between them. *No association with psoriasis susceptibility* is another example of this kind. This is not an error made by the annotator itself but extra work needs to be done to handle the relation between gene the disease. A simple solution would be to ignore text that contains a predefined set of keywords indicating negative relationships, for example, 'not','unlikely','without','no','infrequent','low' and 'unaffected'. A more reasonable solution would be to consider the level of certainty of the relation and extract such

information implied in the text with relation extraction (RE) methods similar to the approach by Bravo et.al [45].

Summary Errors in the *HDGDB* arise from a variety of sources. Some errors are inevitable in an automated process like that used by *OntoSuite-Miner*, but some can be avoided or corrected. These errors have different impacts based on how the data is used. In the case of enrichment analysis discussed in chapter 3 on page 113, because the errors affect both the set of interest and the reference set equally, the errors will most likely cancel each other out when computing statistical enrichment, though this is not guaranteed. In other cases like studying relationships between diseases, these errors may bring in false positive links which may have negative effects. Fortunately, the succinctness and accuracy of the curated databases used as sources for the *HDGDB* mean that annotation errors occur less frequently than, for example, using abstracts or full paper text as sources, which may be more complex and error prone since, for example, diseases may be mentioned but not directly related to the gene under study.

2.3.4.2 Validation of HDGDB against an OMIM 'gold standard' dataset

Ontology terms are usually created with external cross references to other ontologies or resources. The HDO term 'DOID:106052 Alzheimer's disease', for example, has in total 36 external cross references including the KEGG database ('KEGG:05010 Alzheimer's disease pathway'), the OMIM database ('OMIM:104300 ALZHEIMER DISEASE; AD') and the ICD10 database ('ICD10CM:G30 Alzheimer's disease with early onset'). These cross references are created manually by domain experts and considered very reliable. The absence of a gold standard HDO annotation makes it hard to evaluate any newly created HDO annotations. Thus, I used the HDO to OMIM cross references and created an annotation dataset referred as 'OMIM-GOLD', which was used as a test dataset to validate the newly generated *HDGDB*.

In HDO, there are 1349 (19.78%) HDO terms that have at least one OMIM reference and in total 2622 (58.5%) out of 4482 unique disorder entries in OMIM database were referenced with at least one gene annotation. OMIM disorders are manually annotated with genes. These genes were transfered to HDO terms via their OMIM cross references (HDO⇆OMIM disorder⇆genes) to form the 'OMIM-GOLD' dataset.

'OMIM-GOLD' contained 1052 HDO term annotated with 2333 unique genes, forming 3222 gene disease associations (GDAs). As shown in fig. 2.23, *OntoSuite-Miner* was able to recover 66.48%, 76.47% and 64.84% of the 'OMIM-GOLD' GDAs

by mining GeneRIF, OMIM and Ensembl variation databases separately. This increases to 90.32% when these data were merged into *HDGDB*. Most of the GDAs were found by both annotators, indicating that they are high confidence discoveries. Even though GDAs identified solely by one annotator are less trustworthy, they contribute to more than 10% of the correct mappings in *HDGDB* when compared to 'OMIM-GOLD'.

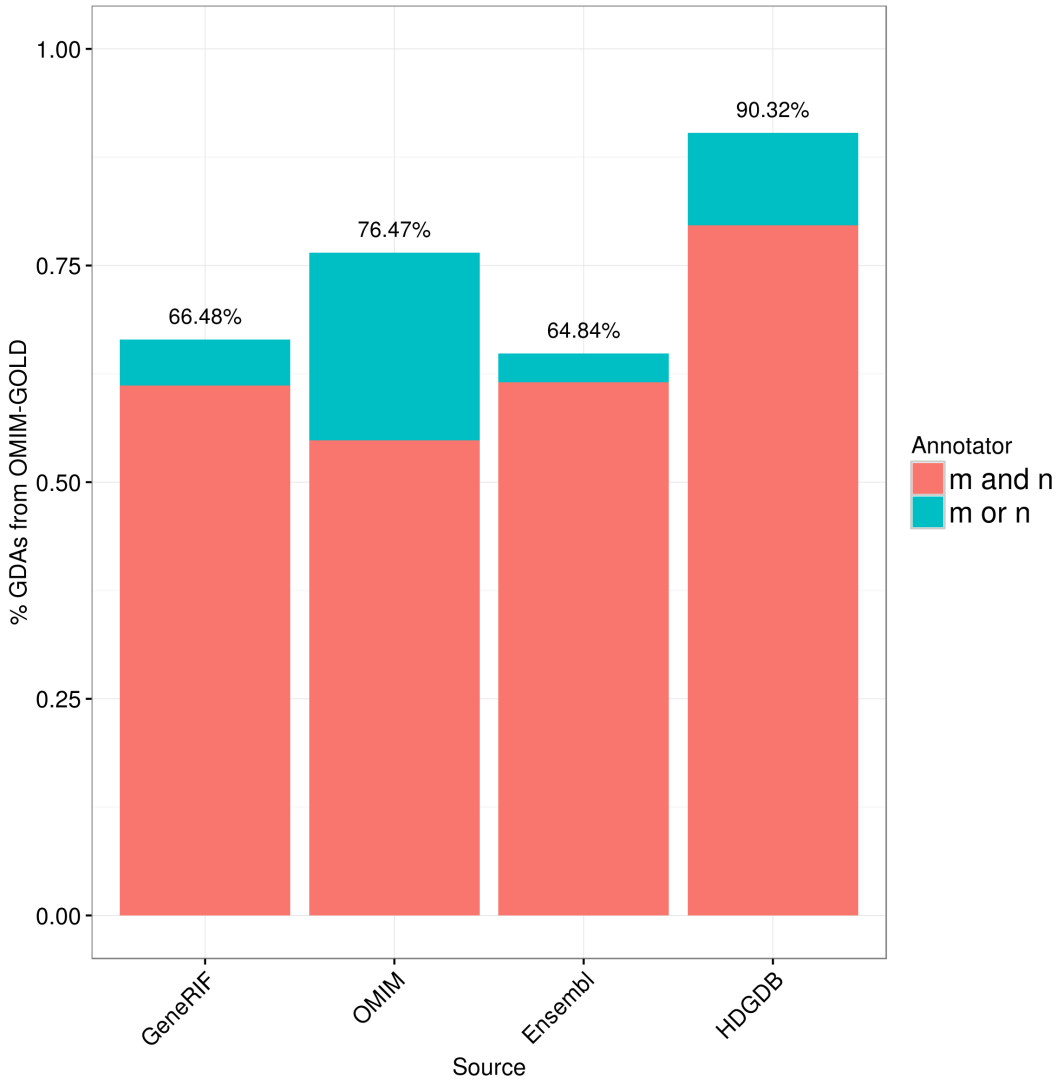


Figure 2.23: The validation of $HDGDB$ by comparing to the 'OMIM-GOLD' HDO annotation dataset (1052 HDO term annotated with 2333 unique genes forming 3222 GDAs) created using OMIM cross-references to HDO terms. $HDGDB$ was able to recover 90.32% of 'OMIM-GOLD' GDAs, indicating a high recall rate while most of the GDAs were found by both annotators, indicating that they are high confidence mappings. Even though GDAs identified solely by one annotator are less trustworthy, they contribute to more than 10% of the correct mappings in $HDGDB$ when compared to 'OMIM-GOLD'. The increased percentages of GDAs recovered from $HDGDB$ compared to the tree individual source highlights the advantages of the data integration.

2.3.4.3 Validation of HDGDB against DisGeNet

Similarly, instead of using OMIM-HDO cross references, as a further validation, I compared *HDGDB* with the DisGeNet [192], which is an integrated knowledge base containing human gene disease association from a variety of sources. However, DisGeNet and *HDGDB* are not directly comparable due to the differences of sources used, but one of the data sources from DisGeNet, the Literature-derived Human Gene-Disease Network (LHGDN) [193], is derived from the GeneRIF database, making it an ideal resource for validating *HDGDB* and the performance of *OntoSuite-Miner* on recovering gene disease association, also from the GeneRIF database.

Recall that in the GeneRIF database, a Pubmed id is indexed with genes and rifs (short sentences submitted by authors that describe the functions of the genes studied in the publication). Each GeneRIF entity thus links one gene to one rif from one publication. LHGDN and *HDGDB*(GeneRIF subset) are both derived by linking such gene/publication pair to disease terms base on its rifs, thus they are comparable. In the following validation process, the LHGDN is considered as a test data set to evaluate *HDGDB*.

LHGDN data was downloaded from DisGeNet website on 28 July 2017. UMLS [10] id was used to represent disease. To enable the comparison, the UMLS ids were mapped to HDO using EBI Ontology Xref Service [161]. Out of the 52828 gene disease associations in LHGDN, 16500(31%) cannot be mapped to HDO, thus removed from further testing. Three categories, *identical*, *better*, and *missed* were used during the assessment. For each gene/publication pair in LHGDN, if *HDGDB*(GeneRIF subset) contains a) the same disease annotation, it is assigned *identical*; b) more specific diseases, it is assigned *better*; c) otherwise, it is assigned *missing*. The result is shown in fig. 2.24.

HDGDB is able to recover in total 30233(88.2%) of the 34288 gene/publication pairs in LHGDN, out of which, 2561(7.47%) contains more specific disease terms. For example, gene *HGFAC*(Entrez gene id 3083) is indexed with Pubmed paper 16189274 [194] with a rif 'HGF/MET signaling and aberrant HGF-activator expression is associated with diffuse large B-cell lymphoma'. LHGDN annotated this gene with 'DOID:707 B-Cell Lymphomas', while *HDGDB* annotated the same gene with 'DOID:0050745 diffuse large b-cell lymphoma', which is a direct child term of the previous in the HDO hierarchy. Another example would be gene *HOXD13* (3239) which is indexed with Pubmed paper 18566322 [195] with a rif 'transgenic mice expressing NUP98-

HOXD13 (NHD13) fusion gene develop myelodysplastic syndrome, and more than half eventually progress to acute leukemia'. In this case, both LHGDN and *HDGDB* corrected annotated the gene with 'DOID:0050908 myelodysplastic syndrome'. However, another disease term 'DOID:1240 leukemia' was annotated to the gene in LHGDN, while in *HDGDB*, one of its direct child term 'DOID:12603 acute leukemia' was found instead.

HDGDB failed to recover 4055(11.8%) annotation for the gene/publication pair compared to LHGDN. However, since LHGDN is a text-mining derived dataset extracted from GeneRIF, it itself contains errors. Thus, there are cases in the 4055 annotation where *HDGDB* correctly identified the annotation while LHGDN did not. The precedent of such cases is however unknown. In another word, the 11.8% missing rate is an upper bound of the error rate while the actual error rate is lower. As an example, gene *CPE*(1363) wan annotated wrongly to 'DOID:169 Neuroendocrine Tumors' in LHGDN base on the rif 'cDNA microarray analysis led to the identification of 2 novel biomarkers that should facilitate molecular diagnosis and further study of pulmonary neuroendocrine tumors' from Pubmed paper 15492986 [196], while *HDGDB* was able to pick up the correct disease term 'DOID:5410 pulmonary neuroendocrine tumor'.

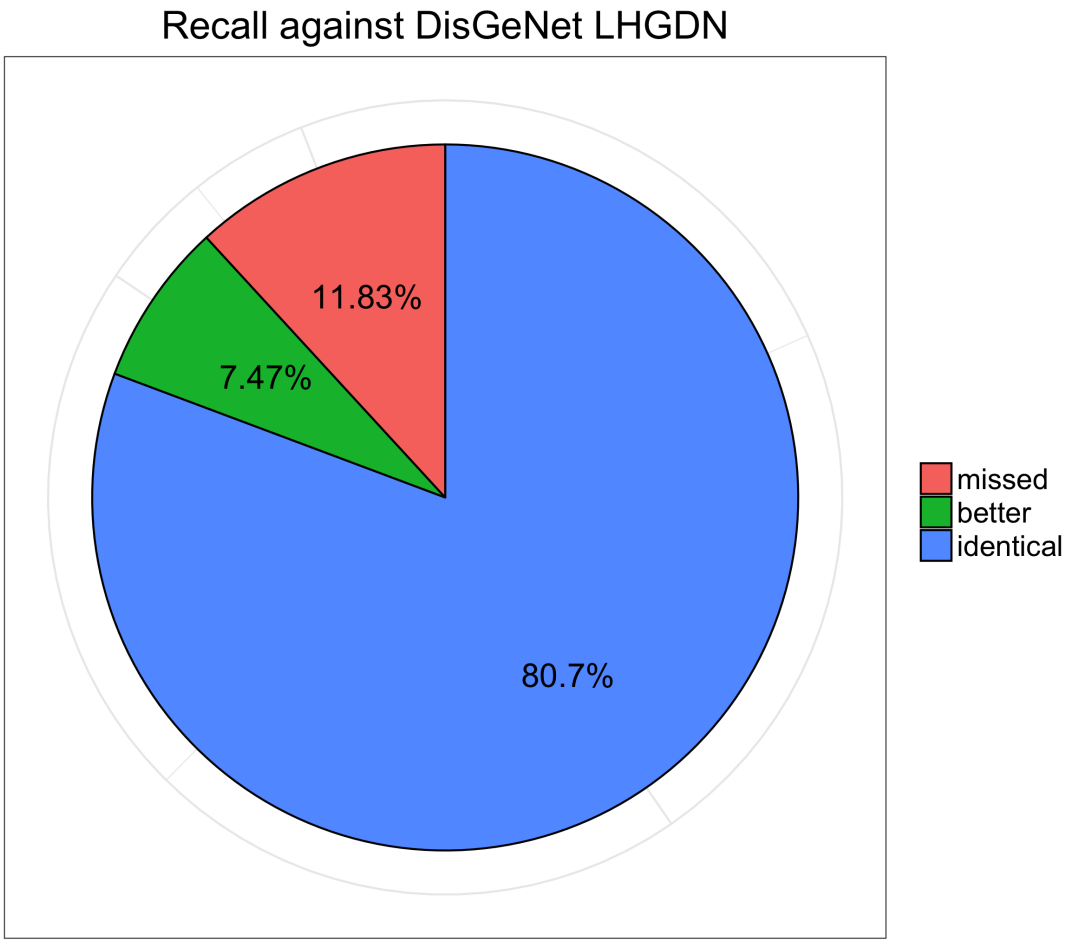


Figure 2.24: The validation of *HDGDB* (GeneRIF subset) against the DisGeNet LHGDN dataset. For each gene/publication pair in LHGDN, it is assigned a) *identical* if *HDGDB*(GeneRIF subset) contains the same disease annotation; b) *better* if *HDGDB*(GeneRIF subset) found more specific diseases, and c) *missing* otherwise. *HDGDB* agrees 80.7% of the GDAs in LHGDN, indicating a high recall rate. 7.47% of the gene/publication pair were annotated with more specific HDO terms in *HDGDB*, while 11.83% were annotated with different HDO terms.

2.3.4.4 Validation of HDGDB against genes from GenAge

I used a set of 305 human genes(build 18, release on October 11, 2015) from the GenAge database [187], which are known to be associated with aging. *HDGDB* was able to annotate 303 (99.34%) of these genes with human disease ontology terms. For

this subset, I ran enrichment analysis with the *topOnto* R package (see chapter 3 on page 113) and obtained the top 50 enriched HDO terms (table 2.10). Most of the enriched diseases are different types of cancer. This agrees with the date published by Cancer Research UK [197] that peak rate of cancer cases, 2012-2014, UK was in people aged 85+ while half of all cancer cases in the UK are diagnosed in people aged 70 and over. Other diseases such as 'Alzheimer's disease', 'type 2 diabetes mellitus' and 'Parkinson's disease' are also known to increase with age.

	TERM.ID	TERM.NAME	Annotated	Significant	Expected	elim-p	elim-p-BY
1	DOID:1909	melanoma	1248	142	21.33	1e-30	5.69e-28
2	DOID:684	hepatocellular carcinoma	1951	167	33.34	1e-30	5.69e-28
3	DOID:1749	squamous cell carcinoma	1518	153	25.94	1e-30	5.69e-28
4	DOID:10283	prostate cancer	2261	190	38.64	1e-30	5.69e-28
5	DOID:8567	Hodgkin's lymphoma	519	85	8.87	1e-30	5.69e-28
6	DOID:299	adenocarcinoma	1465	142	25.03	1e-30	5.69e-28
7	DOID:10652	Alzheimer's disease	1511	122	25.82	1e-30	5.69e-28
8	DOID:0050865	tongue squamous cell carcinoma	636	84	10.87	1e-30	5.69e-28
9	DOID:769	neuroblastoma	620	83	10.6	1e-30	5.69e-28
10	DOID:768	retinoblastoma	306	65	5.23	1e-30	5.69e-28
11	DOID:5520	head and neck squamous cell carcinoma	336	63	5.74	1e-30	5.69e-28
12	DOID:1036	chronic leukemia	465	71	7.95	1e-30	5.69e-28
13	DOID:8552	chronic myeloid leukemia	364	64	6.22	1e-30	5.69e-28
14	DOID:9538	multiple myeloma	534	74	9.13	1e-30	5.69e-28
15	DOID:2237	hepatitis	950	95	16.23	1e-30	5.69e-28
16	DOID:657	adenoma	575	76	9.83	1e-30	5.69e-28
17	DOID:9261	nasopharynx carcinoma	436	66	7.45	1e-30	5.69e-28
18	DOID:2043	hepatitis B	460	67	7.86	1e-30	5.69e-28
19	DOID:1712	aortic valve stenosis	353	60	6.03	1e-30	5.69e-28
20	DOID:1936	atherosclerosis	565	72	9.66	1e-30	5.69e-28
21	DOID:289	endometriosis	458	71	7.83	1e-30	5.69e-28
22	DOID:1612	breast cancer	3487	229	59.59	1e-30	5.69e-28
23	DOID:9119	acute myeloid leukemia	674	86	11.52	1e-30	5.69e-28
24	DOID:3070	malignant glioma	1357	146	23.19	1e-30	5.69e-28
25	DOID:3347	osteosarcoma	522	71	8.92	1e-30	5.69e-28
26	DOID:3910	lung adenocarcinoma	485	66	8.29	1e-30	5.69e-28
27	DOID:4362	cervical cancer	702	94	12	1e-30	5.69e-28
28	DOID:3717	gastric adenocarcinoma	243	50	4.15	1e-30	5.69e-28
29	DOID:0050908	myelodysplastic syndrome	231	48	3.95	1e-30	5.69e-28
30	DOID:9952	acute lymphocytic leukemia	605	69	10.34	1e-30	5.69e-28
31	DOID:1115	sarcoma	463	72	7.91	1e-30	5.69e-28
32	DOID:11054	urinary bladder cancer	796	99	13.6	1e-30	5.69e-28
33	DOID:3068	glioblastoma multiforme	359	55	6.13	1e-30	5.69e-28
34	DOID:3008	invasive ductal carcinoma	376	56	6.43	1e-30	5.69e-28
35	DOID:0050866	oral squamous cell carcinoma	504	63	8.61	1e-30	5.69e-28
36	DOID:219	colon cancer	1137	118	19.43	1e-30	5.69e-28
37	DOID:14221	metabolic syndrome X	425	57	7.26	1e-30	5.69e-28
38	DOID:0050745	diffuse large B-cell lymphoma	254	46	4.34	1e-30	5.69e-28
39	DOID:10286	prostate carcinoma	220	49	3.76	1e-30	5.69e-28

Continued on next page

Table 2.10 – continued from previous page							
	TERM.ID	TERM.NAME	Annotated	Significant	Expected	elim-p	elim-p-BY
40	DOID:10236	exhibitionism	209	42	3.57	1e-30	5.69e-28
41	DOID:2394	ovarian cancer	1486	148	25.39	1e-30	5.69e-28
42	DOID:11714	gestational diabetes	147	37	2.51	1e-30	5.69e-28
43	DOID:9352	type 2 diabetes mellitus	298	48	5.09	1e-30	5.69e-28
44	DOID:3770	pulmonary fibrosis	268	46	4.58	1e-30	5.69e-28
45	DOID:4001	ovarian carcinoma	470	74	8.03	1e-30	5.69e-28
46	DOID:12704	ataxia telangiectasia	78	29	1.33	1e-30	5.69e-28
47	DOID:326	ischemia	388	66	6.63	1e-30	5.69e-28
48	DOID:3908	non-small cell lung carcinoma	1117	126	19.09	1e-30	5.69e-28
49	DOID:12603	acute leukemia	271	45	4.63	1e-30	5.69e-28
50	DOID:643	progressive multifocal leukoencephalopathy	138	35	2.36	1e-30	5.69e-28

Table 2.10: Enrichment analysis for Aging-related genes from the GenAge database *HDGDB* and *topOnto* R package (see chapter 3 on page 113). The enriched disease make biological sense. Most of these disease such as different types of cancer, neurodegenerative disease such as 'Alzheimers disease' and 'Parkinsons disease' are diseases primarily associated with older people with incidence rates increasing with age.

2.3.4.5 Validation of HDGDB against genes from Cildb

Cilia are tiny hair-like organelles formed on the surface of cells and are present on almost all polarized cell types of the human body and are involved in various cellular functions. Cilia can either be motile or immotile (sometimes refered to as sensory cilia or primary cilia). Ciliary motility is required to move extracellular fluid while immotile cilia are thought to have sensory and signaling roles. Ciliary dysfunction causes a number of diseases in humans from development defects to defects in vision, smell, and hearing. Well known cilia-related diseases includes 'primary ciliary dyskinesia (PCD)', 'situs inversus totalis' and 'Nephronophthisis' are referred to as ciliopathies, and the number of diseases caused by ciliary dysfunction is expected to increase (see review paper in [198, 199]).

As part of the validation of the *HDGDB* data generated by *OntoSuite-Miner*, I first explored *HDGDB* for known ciliopathies to ensure that the data recapitulated known disease associations. Six well known ciliopathies were picked and the number of genes annotated to each disease from each source were calculated (Table 2.11). Some GDAs were identified by multiple sources while others were only found in a single source. By integrating such information into *HDGDB*, the inter-connection between these dis-

eases was revealed (Figure 2.25). For example, 'primary ciliary dyskinesia' shares no genes between any of the diseases in OMIM , GeneRIF or Ensembl variation database but is linked to 'polycystic kidney disease' by gene *CCDC151* in *HDGDB*.

	OMIM	GeneRIF	Ensembl variation	*HDGDB*
primary ciliary dyskinesia	26	22	27	56(17)
Joubert syndrome	23	2	68	75(23)
Bardet-Biedl syndrome	21	23	72	76(23)
Senior-Loken syndrome	7	3	13	16(5)
polycystic kidney disease	3	47	14	57(17)
polycystic liver disease	2	5	6	9(3)

Table 2.11: Number of genes annotated to six well known ciliopathies. Some genes were identified by multiple sources while others were only found in a single source. The number in the bracket indicates the number of false positives estimated base on the result (Precision rate 69.84% on GWAS_GeneRIF(HDC)) from section 2.3.1.

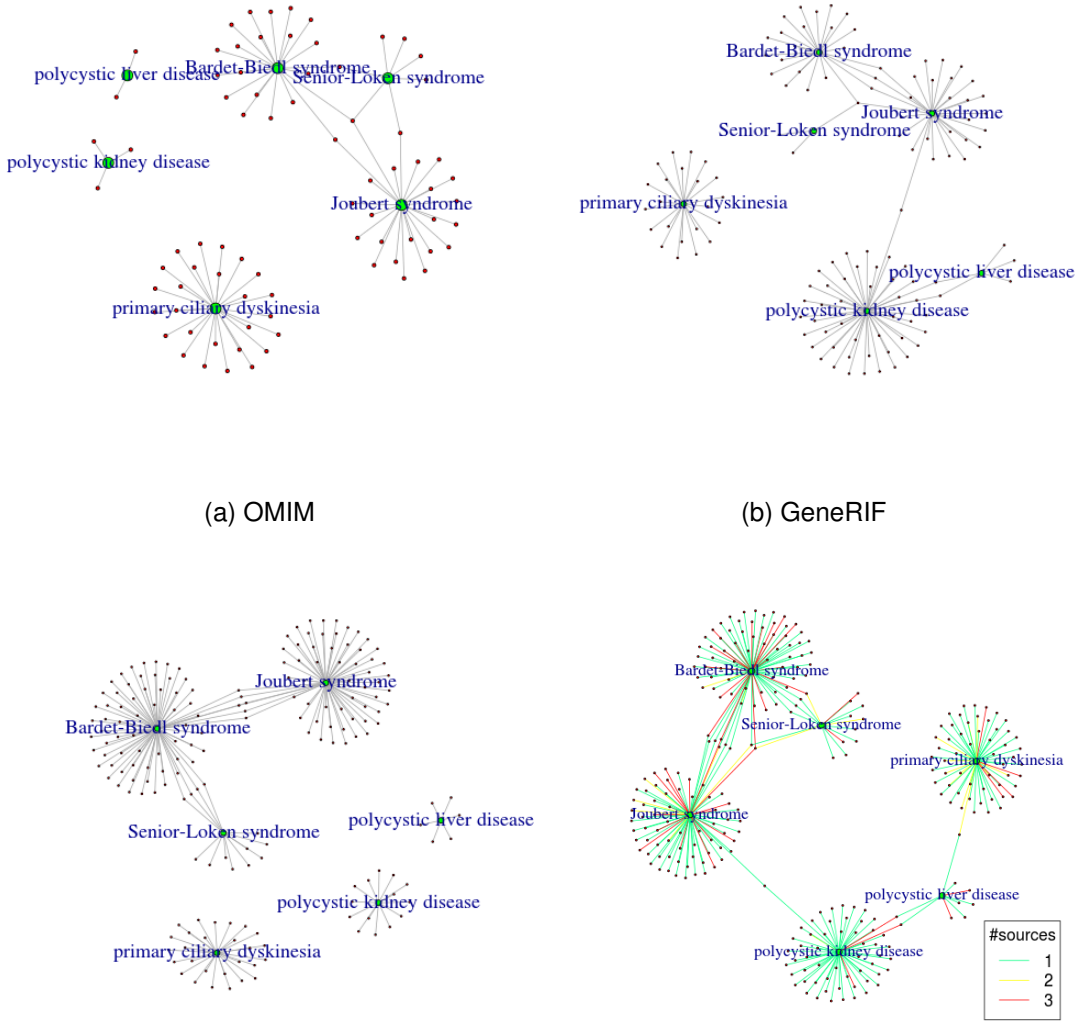
As a further validation of the utility of the *HDGDB*, I performed disease enrichment analysis with *topOnto* R package (discussed in the chapter 3 on page 113) on a set of cilia-related genes from Cildb [188]. The gene data, comprised 3133 human genes having at least one piece of evidence of association with cilia and was taken from from Cildb on 25/04/2016. 446 diseases were enriched with $p \leq 0.05$. For readability, the top 50 significant enriched terms are shown in table 2.12 while the full list can be found as supplementary information in Cildb_enrichment.csv (attached disk).

	TERM.ID	TERM.NAME	Level	Annotated	Significant	elim-p	elim-p-BY
1	DOID:684	hepatocellular carcinoma	9	1951	406	4.0e-20	1.21e-15
2	DOID:9562	primary ciliary dyskinesia	5	56	36	3.3e-18	4.98e-14
3	DOID:1749	squamous cell carcinoma	7	1518	326	3.4e-17	3.42e-13
4	DOID:9261	nasopharynx carcinoma	8	436	118	8.2e-14	5.53e-10
5	DOID:1909	melanoma	6	1248	263	9.8e-14	5.53e-10
6	DOID:1612	breast cancer	7	3487	671	1.1e-13	5.53e-10
7	DOID:219	colon cancer	10	1137	255	3.3e-12	1.42e-08
8	DOID:0050865	tongue squamous cell carcinoma	9	636	150	6.4e-12	2.41e-08
9	DOID:1324	lung cancer	7	2427	474	8.9e-11	2.98e-07
10	DOID:299	adenocarcinoma	7	1465	300	1.7e-10	5.13e-07
11	DOID:3008	invasive ductal carcinoma	10	376	97	2.5e-10	6.54e-07
12	DOID:3498	pancreatic ductal adenocarcinoma	10	304	83	2.6e-10	6.54e-07
13	DOID:1107	esophageal carcinoma	8	789	179	3.1e-10	7.19e-07
14	DOID:3969	papillary thyroid carcinoma	9	244	70	6.4e-10	1.38e-06
15	DOID:0050866	oral squamous cell carcinoma	8	504	119	9.5e-10	1.91e-06

Continued on next page

Table 2.12 – continued from previous page							
	TERM.ID	TERM.NAME	Level	Annotated	Significant	elim-p	elim-p-BY
16	DOID:768	retinoblastoma	11	306	81	2.1e-09	3.96e-06
17	DOID:10283	prostate cancer	8	2261	406	1.8e-08	3.19e-05
18	DOID:4001	ovarian carcinoma	11	470	122	2.5e-08	4.19e-05
19	DOID:8469	influenza	5	240	65	3.1e-08	4.83e-05
20	DOID:10619	lymph node cancer	8	205	58	3.2e-08	4.83e-05
21	DOID:2876	laryngeal squamous cell carcinoma	9	194	55	6.6e-08	9.48e-05
22	DOID:1115	sarcoma	6	463	103	3.0e-07	4.11e-04
23	DOID:305	carcinoma	6	4065	768	3.4e-07	4.46e-04
24	DOID:3910	lung adenocarcinoma	10	485	106	5.1e-07	6.41e-04
25	DOID:3565	meningioma	8	152	44	7.0e-07	8.45e-04
26	DOID:3068	glioblastoma multiforme	8	359	83	8.2e-07	9.51e-04
27	DOID:8567	Hodgkin's lymphoma	9	519	114	9.3e-07	1.04e-03
28	DOID:3113	papillary carcinoma	7	67	25	1.2e-06	1.29e-03
29	DOID:11054	urinary bladder cancer	7	796	173	1.6e-06	1.66e-03
30	DOID:0050639	primary cutaneous amyloidosis	6	272	66	1.9e-06	1.91e-03
31	DOID:2394	ovarian cancer	8	1486	309	2.2e-06	2.14e-03
32	DOID:3114	serous cystadenocarcinoma	9	145	41	3.2e-06	3.02e-03
33	DOID:0050777	Joubert syndrome	5	75	26	3.6e-06	3.29e-03
34	DOID:10459	common cold	6	156	43	3.9e-06	3.46e-03
35	DOID:4450	renal cell carcinoma	9	759	148	4.1e-06	3.53e-03
36	DOID:1712	aortic valve stenosis	8	353	79	5.6e-06	4.57e-03
37	DOID:657	adenoma	6	575	117	5.6e-06	4.57e-03
38	DOID:1793	pancreatic cancer	7	1392	283	6.0e-06	4.76e-03
39	DOID:0050569	Seckel syndrome	7	32	15	6.2e-06	4.80e-03
40	DOID:3144	cutis laxa	6	29	14	8.3e-06	6.25e-03
41	DOID:0050576	Senior-Loken syndrome	7	16	10	8.5e-06	6.25e-03
42	DOID:8552	chronic myeloid leukemia	10	364	80	9.9e-06	7.11e-03
43	DOID:1485	cystic fibrosis	7	269	63	1.1e-05	7.54e-03
44	DOID:0050908	myelodysplastic syndrome	9	231	56	1.1e-05	7.54e-03
45	DOID:2871	endometrial carcinoma	10	343	76	1.2e-05	8.04e-03
46	DOID:1935	Bardet-Biedl syndrome	7	76	25	1.6e-05	1.05e-02
47	DOID:9538	multiple myeloma	10	534	108	1.7e-05	1.09e-02
48	DOID:0050902	medulloblastoma	10	240	57	1.8e-05	1.13e-02
49	DOID:769	neuroblastoma	9	620	122	1.9e-05	1.15e-02
50	DOID:4610	intestinal benign neoplasm	7	228	58	1.9e-05	1.15e-02

Table 2.12: Enriched diseases resulting from an enrichment analysis using topOnto and the *HDGDB* for a list of 3133 human genes from CilDB. f'Level' indicates the depth of the disease term in the Human Disease Ontology hierarchical structure. The total number of gene annotated to a term in *HDGDB* and the number of overlapping genes in the CilDB gene list is shown in the 'Annotated' column and the 'Significant' column. A number of different types of cancer including lung/breast/prostate/pancreatic/skin carcinoma, lymphoma, leukemia and sarcoma were significantly enriched. This result suggests that cilia may play an important role under the common theme of cancer where each different type of cancer shares a subset of overlapping developing mechanism which require the involvement of cilia.



(a) OMIM (b) GeneRIF

(c) Ensembl variation (d) HDGDB

Figure 2.25: Known ciliopathies and their annotated genes in OMIM,GeneRIF,Ensembl variation and *HDGDB*. Green nodes represent diseases while red nodes represent genes. Genes were connected to diseases by edges based on their sources. (a) and (b) shows the gene disease network built on GDAs from the OMIM and GeneRIF databases. The networks are relatively small and there are not many connections between diseases. (c) shown the network built using the Ensembl Variation database which consists mostly of data from GWAS studies. GWAS is able to pick up a lot of potential genes for diseases, thus have a better recall but bigger noise. As a result, many more genes were identified from the Ensembl Variation database but some diseases are still orphan. (d) The network built using *HDGDB*. All of the diseases are now connected by at least one gene. Edge color indicates the number of sources contributing to the annotation. Interestingly, 'primary ciliary dyskinesia' shares no genes with any of the diseases in OMIM, GeneRIF or Ensembl variation database but is linked to 'polycystic kidney disease' by gene *CCDC151* in *HDGDB*.

Interestingly, on examining the list of enriched diseases, I found that a number of different types of cancer including lung/breast/prostate/pancreas/skin carcinoma, lymphoma, leukemia and sarcoma were significantly enriched. This result suggests that cilia may play an important role in cancer-related processes where each different type of cancer shares a subset of overlapping mechanisms which involve of cilia. In fact, the relationship between cilia and cancer is an emerging research area (see a recent review by Cao et. al. [200]). Cilia are required to promote the formation of the spindle during mitosis thus play a role in control of cell division, cell polarity and cell migration. They can suppress abnormal cell growth and proliferation by affecting the cell cycle. Dysfunction of cilia has been proposed as a prerequisite step for cancer development [201] and has been observed in multiple cancer types [202]. Despite a much-improved understanding of the relationship between cilia and cancer relation and association, little is known of their direct role in tumorigenesis.

Next, I examined a list of 9 known ciliopathies and 14 suspected ciliopathies from [203] and their enrichment status in the enrichment results from Cildb genes. HDO was not used in [203], thus ciliopathies from [203] were manually mapped to HDO terms based on their names. For example, 'Dandy-Walker malformation' was mapped to the HDO term 'Dandy-Walker syndrome'. 7 out of 9 known ciliopathies were found to be significantly enriched but there was no statistical support for the enrichment of 'Alstrom syndrome' or 'orofaciodigital syndrome I' (table 2.13). Among the 14 suspected ciliopathies, 6 were found to be enriched. Note that 'diabetes mellitus' was not enriched when using OMIM, GeneRIF or Ensembl Variation alone, but was found to be enriched using *HDGDB*, illustrating the advantage of integrating multiple data sources to increase the statistical power of enrichment analysis. What's more, general term 'diabetes mellitus' was enriched but none of its children terms ,for example type I diabetes or type II diabetes, were enriched, indicating indicating that the enrichment was the join result of all children terms of diabetes, which would have been missed without the topology information (see chapter 3 on page 113 for details.) As potentially more data from other sources being integrated into *HDGDB* in the future, the enrichment result of Cildb genes is likely to confirm other suspected ciliopathies or revealing possible diseases that have not been previously linked to cilia dysfunction.

Disease network analysis has emerged as a powerful way of studying inter-connections between diseases [204–206]. Such networks of diseases and disease genes offers a platform to explore all known diseases and disease gene associations in a single graph

	TERM.ID	TERM	OMIM	GeneRIF	Ensembl variation	*HDGDB*
Known Ciliopathies						
1	DOID:0050473	Alstrom syndrome	-	Y	-	-
2	DOID:1935	Bardet-Biedl syndrome	Y	Y	Y	Y
3	DOID:0050777	Joubert syndrome	Y	Y	Y	Y
4	DOID:0050778	Meckel syndrome	Y	Y	Y	Y
5	DOID:12712	nephronophthisis	Y	Y	Y	Y
6	DOID:0060316	orofaciodigital syndrome I	-	-	-	-
7	DOID:0050576	Senior-Loken syndrome	Y	Y	Y	Y
8	DOID:898	polycystic kidney disease	-	Y	-	Y
9	DOID:9562	primary ciliary dyskinesia	Y	Y	Y	Y
Suspected Ciliopathies						
1	DOID:4626	hydranencephaly	-	-	-	-
2	DOID:2785	Dandy-Walker syndrome	-	-	-	-
3	DOID:9351	diabetes mellitus	-	-	-	Y
4	DOID:12714	Ellis-Van Creveld syndrome	-	-	-	-
5	DOID:409	liver disease	-	-	-	-
6	DOID:0050592	asphyxiating thoracic dystrophy	-	Y	-	Y
7	DOID:0060172	juvenile absence epilepsy	-	-	-	-
8	DOID:9970	obesity	-	-	-	-
9	DOID:1148	polydactyly	Y	Y	-	Y
10	DOID:11162	respiratory failure	-	-	-	-
11	DOID:10003	sensorineural hearing loss	-	-	-	Y
12	DOID:758	situs inversus	Y	-	Y	Y
13	DOID:10584	retinitis pigmentosa	-	Y	-	Y
14	DOID:0080016	spina bifida	-	-	-	-

Table 2.13: Enrichment analysis for known and suspected Ciliopathies taken from [203]. Enrichment analysis was carried out with *TopOnto* and the p-value was calculated using Fisher's exact test with the *elim* method. A term is considered to be enriched with $p \leq 0.05$. Seven out of 9 known ciliopathies were found to be significantly enriched but there was no statistical support for the enrichment of 'Alstrom syndrome' and 'orofaciodigital syndrome I'. Among the 14 suspected ciliopathies, 6 were found to be enriched. Note that 'diabetes mellitus' was not enriched when using OMIM, GeneRIF or Ensembl Variation alone but was enriched with $HDGDB$, indicting the advantage of integrating multiple data sources to increase the statistical power of enrichment analysis.

theoretic framework, and has already been used to reveal the common genetic origin of many diseases as well as predict potential disease gene candidates [204, 206, 207]. In order to explore the inter-connections of ciliopathies, a disease network was constructed with the top enriched diseases from the Cildb enrichment analysis (fig. 2.27, to reduce the complicity and increase the visibility as a graph in the thesis, only the top 100 enriched diseases were used). Nodes in the network represent diseases. Two nodes are connected by a weighted edge if they share CilDB gene(s) with the weight of the edge equal to the number of common genes. With the rationale that ciliopathies are a collection of related diseases, they are likely to share similar molecular mechanisms and group together in the same community in the network. Thus, the community detection algorithm *spinglass.community* was applied to the network to identify ciliopathy communities. A community is defined as a set of nodes with many edges inside the community and few edges outside it (i.e. between the community itself and the rest of the network). The community which contains the largest number of known ciliopathies is referred to as a ciliopathy community. In total, 9 known ciliopathies (table 2.14) were picked as markers by manually inspecting the enriched diseases in the Clidb genes.

TERM.ID	TERM
DOID:0060340	ciliopathy
DOID:1485	cystic fibrosis
DOID:2975	cystic kidney
DOID:0050777	Joubert syndrome
DOID:10584	retinitis pigmentosa
DOID:0050576	Senior-Loken syndrome
DOID:1935	Bardet-Biedl syndrome
DOID:9562	primary ciliary dyskinesia
DOID:3083	chronic obstructive pulmonary disease

Table 2.14: 9 known ciliopathies that were enriched in the Cildb genes.

An argument, *gamma*, is defined to control the community detection algorithm, specifying the balance between the importance of edges in a community. The default value of 1.0 makes existing and non-existing edges equally important. Roughly, small *gamma* values generate a few big communities while larger *gamma* values generate more communities (fig. 2.26).

To determine the appropriate *gamma* value for the disease network, I define $n(\gamma)$ the number of community generated for a give γ, $N_i(\gamma)$ the number of known ciliopathies in the i_{th} community, $S_i(\gamma)$ the size of the i_{th} community. Let j be the index of the ciliopathy community for a give γ, so N_j is the number of ciliopathy disease in the ciliopathy community while S_j is the size of the ciliopathy community. The Objective function is written as follows:

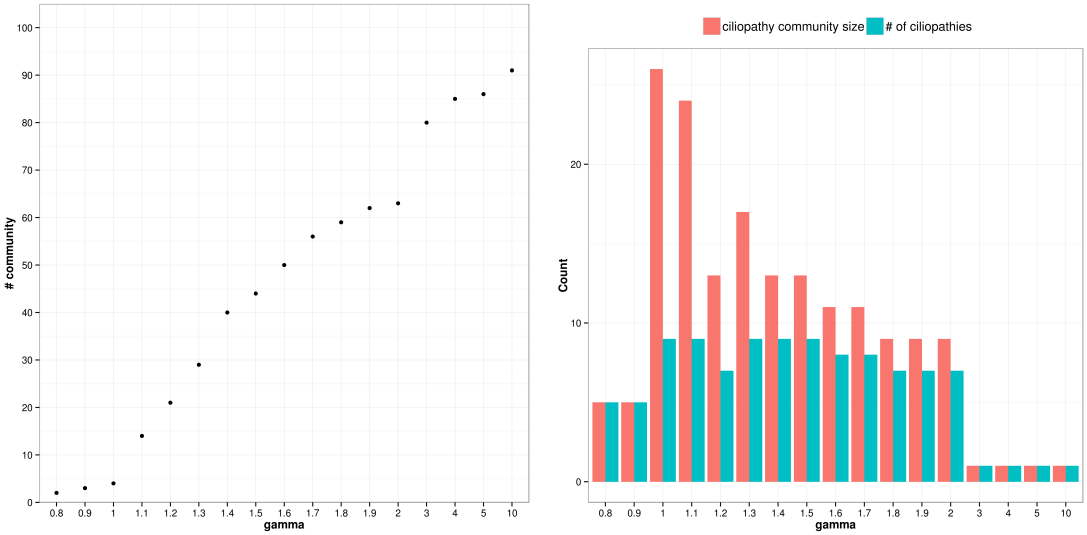
$$\max_{\gamma} N_j(\gamma) \tag{2.8}$$

$$\text{s.t. } S_j(\gamma) = \max(S_1(\gamma) \dots S_n(\gamma)) \tag{2.9}$$

$$\text{where } N_j(\gamma) = \max(N_1(\gamma) \dots N_n(\gamma)) \tag{2.10}$$

To determine the appropriate *gamma* value for the disease network, the algorithm was applied to the network with a range of predefined *gamma* values in order to find the one that generates a maximum N_j(eq. (2.9)) subjecting to eq. (2.9). The resulting ciliopathy community size and the number of known ciliopathies in that community is shown in fig. 2.26b. The best result (referred to as the ciliopathy community) was generated with a *gamma* value equal to 1.3, and comprised 17 diseases, including all 9 known ciliopathies, forming the largest community in the network (fig. 2.27). Note that when *gamma* value equal to 1, the generated ciliopathies community size is larger than the one when *gamma* value equal to 1.3, but this ciliopathies community is not the largest community in the network, i.e. there were other larger non-ciliopathies communities in the network.

In addition to the 9 known ciliopathies, the ciliopathy community also contained 8 novel enriched diseases. Some of these diseases are known to involve cilia function while others have either little or no association with cilia. Respiratory system disease like common cold was enriched and fell into the ciliopathy community. In human, cold viruses are transported by nasal cilia to the front of the nasal passages where they infect nasal cells. Loss of cilia and ciliated cells was observed and suggested to be responsible for the impaired mucociliary function in patients with common cold [208]. Numerous cilia-related diseases have been described in [198] that are associated with developmental defects or degeneration affecting the central nervous system which explain the enriched diseases including 'cerebral degeneration', 'Alzheimer's disease', 'frontotemporal dementia' in the ciliopathy community. The occurrence of ischemia and Diamond-Blackfan anemia (DBA) in the community is very interesting. Same evidence has shown that the change of functional cilia length/mass is associated with

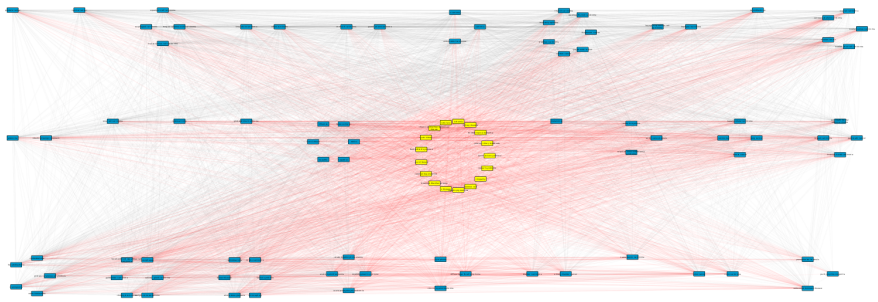


(a) (b)

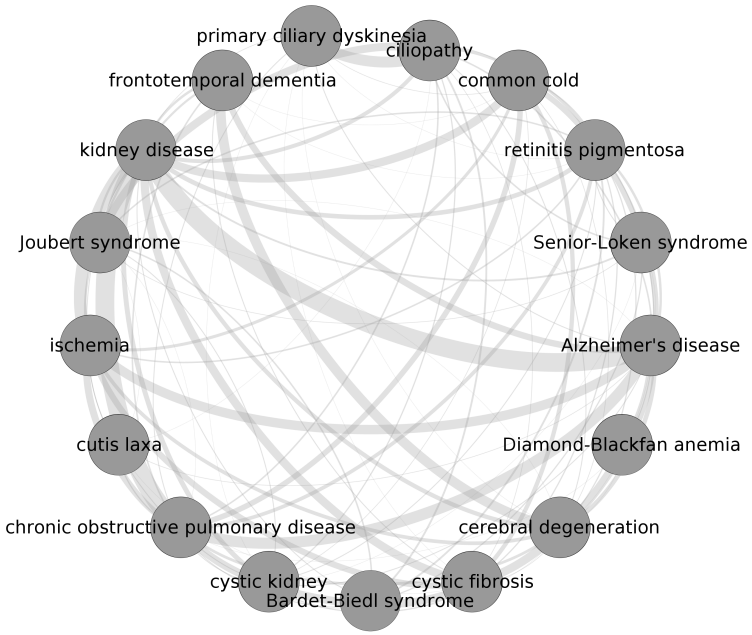
Figure 2.26: Testing different *gamma* values for community detection applied to the ciliopathy disease network. Small *gamma* values generate a few big communities while larger *gamma* values generate more smaller communities. (a) *gamma* affects the number of communities in the network and (b) network and ciliopathy community (the largest number of known ciliopathies) size under different *gamma* settings.

ischemia [209, 210] but to the best of this author's knowledge, there is no previous evidence supporting the association of DBA to cilia. DBA is a congenital erythroid aplasia that usually presents in infancy, causing low red blood cell counts (anemia) [211]. Its gene signature is similar to some of the known ciliopathies suggesting a common underlying molecular mechanism but its connection with cilia function is still unknown and requires further investigation.

Another interesting member found in the ciliopathy community is cutis laxa, a group of rare connective tissue disorders in which the skin becomes inelastic and hangs loosely in folds. The loose skin is often most noticeable on the face but can also affect internal organs including lungs and heart when it is severe. It has not been classified as a ciliopathy. A search in the Pubmed database with keyword 'cutis laxa' and 'ciliopathy' returned only 5 recent papers, all of which reported evidence of having cutis laxa as a comorbidity of other cilia-related diseases. For example, Alazami et al. reported a very rare autosomal recessive disorder, Cranioectodermal dysplasia (CED), characterized by a recognizable craniofacial profile in addition to ectodermal manifestations involving the skin, hair, and teeth [212]. It is highly likely that CED is a ciliopathy



(a)



(b)

Figure 2.27: (a) a disease network comprising the top 100 enriched diseases from enrichment analysis of a list of cilia-related genes from CilDB. Nodes in the network represent diseases. Two nodes were connected by weighted edge if they shared CilDB gene(s), the weight of the edge is equal to the number of common genes. The network community was generated using the *spinglass.community* algorithm from *igraph* with a *gamma* value of 1.3. The ciliopathy community (yellow) community is the largest community in the network with 17 members containing all 9 master ciliopathy diseases. (b) Diseases in the ciliopathy community. The width of the edge represent the weight (genes in common between two nodes). High resolution gene disease network of the ciliopathy community available in the attached disk.

because four genes involved in the ciliary intraflagellar transport were known to be mutated in this disorder. In the report, typical CED features were observed in a multiplex consanguineous family in addition to intellectual disability and markedly lax skin with joint laxity fulfilling the clinical definition of cutis laxa. However, none of the studies had linked cutis laxa directly to cilia dysfunction. According to the disease annotation data from HDGDB, cutis laxa was annotated with genes including ALDH18A1, ELN and FBLN5 which have been found associated with multiple known ciliopathies including chronic obstructive pulmonary disease, cystic fibrosis and joubert syndrome [213–216]. Such overlapping suggested that cutis laxa may be another undiscovered family member of ciliopathies.

Summary Cilia are located on almost all polarized cell types of the human body. The malfunction of cilia can result in a number of human disorders and the list is expected to grow. By running enrichment analysis with cilia-related genes with our automatic generated disease annotation database, I demonstrated that *HDGDB* is able to capture most of the well known ciliopathies as well as novel diseases that have not yet been connected with ciliopathies. Our results suggest that abnormal function of the cilia plays an important role in diseases like cutis laxa and provides insight into the molecular mechanism of these diseases.

2.3.5 Extending annotation to model species

Model organisms are widely used to understand biological phenomena. Functional information can often be transfered between human genes a genes in model organisms so that discoveries made in the model organism can provide insight into the workings of the human diseases. A possible scenario would be the identification of a set of differentially expressed genes between pathological and control states, e.g., disease vs. healthy phenotypes. These genes may then be mapped from the model species to their homolog genes in the human to reveal information relevant to human health and disease.

Homologous relationships are used to follow gene functions between different species, however, not all the homolog can be used in this way. As shown in fig. 2.28, there are two main homology types: Orthologues and Paralogues.

Orthologs are genes in different species that have evolved from a common ancestral gene by specification. Paralogs are genes related by duplication within a genome

which can happen in the same species or across different species. Orthologs are genes in different species that descend by speciation from the same gene in the last common ancestor [217]. Orthologs are generally assumed to retain equivalent functions in different organisms and to share other key properties even if they have diverged since the speciation event [218]. While orthologs generally retain the same function through the course of evolution, paralogs often evolve new functions, so in the case of transfer, functional information between human and model organisms, orthologs are considered more suitable.

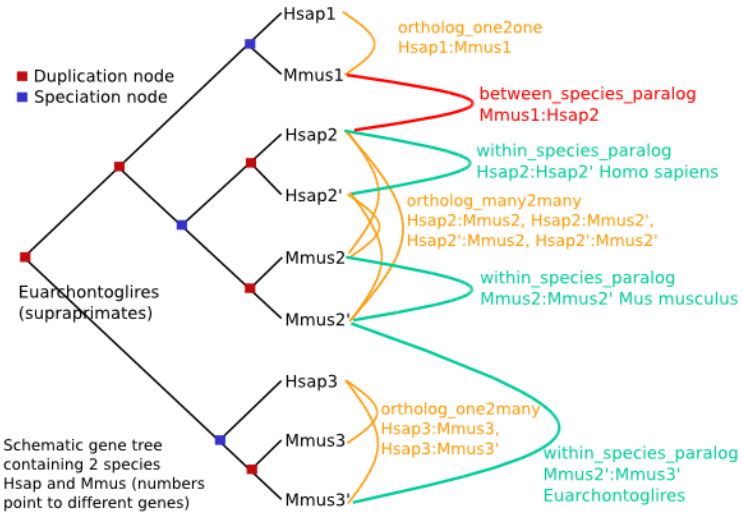


Figure 2.28: Different homology types [219]. Genes in different species and related by a speciation event are defined as orthologs. 1:1, 1:many and many:many relationships are defined depending on the number of genes found in each species. Genes of the same species and related by a duplication event are defined as paralogs.

Orthologs can be further classified into one-to-one orthologues, one-to-many orthologues and many-to-many orthologues. When mapping human genes to model organisms, for example Drosophila melanogaster, a one-to-one orthology relationship means that a human gene has only one ortholog in Drosophila which is a good indication that the gene function is likely to be conserved and can be transferred. A one-to-many relation means a single human gene is orthologous to several Drosophila genes. In this case it is very likely that one or several of the Drosophila genes shares the same function (or part of it) as the human genes (the fewer in-paralogs the more likely they share the function in Drosophila). These annotations could also be transferred with the cautionary note that they are based on one-to-many relationships. Functional transfers with many-to-one and many-to-many orthology relationships are less certain but can

still be informative. *HDGDB* contains integrated disease annotation for human genes. Such information can be transfered to different model organisms. In order to extend the use of such annotations, I mapped these annotations to different model species using homology data from NCBI HomoloGene [123]. Many-to-one and many-to-many orthology relationships were removed to keep the annotation as accurate as possible. A pipeline (fig. 2.29) was implemented to automate the mapping and updating process due to the fact that the homology data are continually changing and evolving in a fast speed.

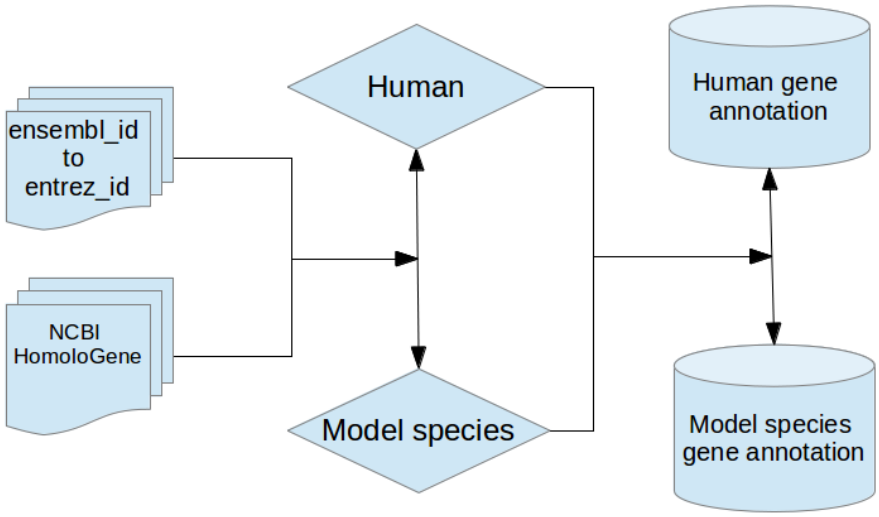


Figure 2.29: Pipeline implemented in *OntoSuite-Miner* to transfer ontology based annotation across species. EntrzGene id is used as the primary index for gene. Ontology base annotation are projected from human to model species with ortholog data collected form the NCBI HomoloGene database.

2.3.5.1 Drosophila melanogaster

Drosophila is a model organism that has been utilized to model a number of human diseases. For example, Drosophila have been used to model and study a wide variety of brain diseases [220]; heart disease [221]; various muscular dystrophies [222–224]; multi-symptom inherited disorders [225]; responses to infection by human pathogens [226–228] and cancer [64, 229]. Drosophila is also used as model of asthma [230], lipotoxicity [231] and metabolism [232, 233]. The conservation of gene function between human and Drosophila makes Drosophila an excellent model to study human

diseases.

To map fly genes to human ones, homolog data were taken from the Ensembl database using Biomart [169] on 10/05/2016. One-to-one and one-to-many orthologs were used to map human genes to fly genes. As a result, 4675 fly genes were annotated with at least one HDO term. Among the 49823 GDAs found in fly, 2977(6%) were scored above 0.3 which indicated that they were from at least two sources. On average, each fly gene was annotated with 10 HDO terms.

This fly gene disease annotation was automatically created and stored in a tab delimited text file. Other model organism GDAs like mouse, rat and zebrafish can easily be added to the pipeline to transfer human gene annotations. When *HDGDB* releases a new update, the fly gene annotation will be automatically updated, as well as annotation data for other future added model species in the pipeline.

2.4 Conclusions and Future work

In this chapter I have presented a methodology to annotate genes by mining biomedical texts for ontology terms of interests. I implemented the method into a framework named *OntoSuite-Miner* with a set of scripting languages including R, Perl and Linux shell command. Most the scripts are available on github (https://github.com/statbio/OntoSuite-Miner) but the two annotators, *MeM* (10GB) and *NcA* (7.4gb) are too big in size, thus stored locally on a Linux server. The framework is highly configurable and customizable, it (a) automatically connects to data source ftp to retrieve up-to-date gene annotation texts, (b) integrates two of the most popular NLP-annotators, MetaMap and the NCBO-annotator, (c) can be configured to auto-run in a preferable time interval to keep track of the latest gene annotations and (d) stores the result in a portable SQLite database allowing easy access within the context of pipeline-based workflows of wider scope. In addition, OntoSuite-Miner also provides an easy to use interface allowing users to submit a text, choose an ontology and get the results ”on the fly”. Moreover, *OntoSuite-Miner* is extensible. Potentially, more annotators,for example, the The ConceptMapper [147], could be included in the framework to improve accuracy and coverage. *OntoSuite-Miner* has been designed to be generalizable so it works uniformly on different ontologies but the performance may vary depending on the ontology used.

I have demonstrated the usage of *OntoSuite-Miner* by mining gene annotation data from three sources including the OMIM, GeneRIF and Ensembl Variation databases for gene-disease associations using the human disease ontology. A gene disease anno-

tation database named *HDGDB* was created and validated quantitatively and qualitatively. Firstly, I analyzed different types of errors in HDGDB. By sampling and manually inspecting errors from 900 annotations from HDGDB, errors were summarized into four types, namely, coordinating conjunctions, abbreviations. Missing concepts/synonyms and others. Coordinating conjunctions were found to be the main sources of error. Possible improvements/solutions were discussed for each type of error in section section 2.3.4.1 on page 76. Secondly, I validated *HDGDB* against a gold standard. This showed that *HDGDB* was able to recover 90.32% of the gold standard annotation a very high recall rate. Finally, I examined the annotations of known aging genes from GenAge database [187] and known cilia-related genes from Cildb [188]. In particular, for the cilia-related gene set, I ran enrichment analysis and looked more closely at a number of subsets of the enriched disease terms and built a network of the inter-connections between diseases based on their shared genes. I utilized a network community algorithm *spinglass.community* and identified a smaller sub-network for which there was a strong biological support for a set of diseases called ciliopathies. The usage of HDGDB, in combination with these analysis and selection techniques, allowed the identification of several novel links that interconnected known ciliopathies in biologically meaningful communities, as well as a number of novel diseases with no previous known link to cilia.

A confidence score have been implemented, indicating the level of certainty of each gene disease association(GDA). The score takes into account the number of sources, the amount of evidence that support the association and the number of annotators that map the association. It provides a way to rank/weight the associations based on the evidence and assists in the prioritization and navigation of the the GDAs. In addition, the score can be applied in analysis such as gene set enrichment analysis (discussed in the next chapter) so that highly ranked GDAs can be preferentially weighted.

Next, I discussed the usefulness and corresponding methodology for extending *HDGDB* to model species. A pipeline was implemented to transfer human gene annotations to model species using one-to-one and one-to-many orthologs. Drosophila melanogaster (fruit fly) which has long been utilized to model a number of human diseases, was used to demonstrate the usage of the pipeline. As a result, 4675 fly gene were annotated with at least one HDO term. On average, each fly gene was annotated with 10 HDO terms. Whenever *HDGDB* releases a new update, the fly data will be update automatically to capture the latest changes. Note that other model organisms can easily be added to the system.

A few points about the implementation and methodology need further discussion. *OntoSuite-Miner* is designed to extract gene associations from plain text. However, the text corpus sometimes describes gene associations, such as'unaffected', which expresses a negative relationship. Gene association can also be unqualified or not specified at the semantic level, e.g. 'Gene A is associated with disease B' or semantically specified, e.g. 'Gene A is over expressed in disease B'. Moreover a gene association may be described with a level of certainty; that is whether the association is phrased as a fact or proven experimental observation or, alternatively, as a hypothesis or speculation. e.g. 'Gene A might be associated with disease B'. The current implementation of *OntoSuite-Miner* does not capture this type of information, thus the resulting data does not provide the causality of gene association. Several approached have been proposed to overcome this problem in the field of relation extraction (RE) including rule-based approaches [133, 234], co-occurrence based statistic approaches [154, 235, 236], machine learning [45, 237–240] and NLP-based systems [241, 242]. Supervised learning approaches have shown good performance identifying relations between entities in text [243]. These approaches usually classify text based on how the relationship is represented [244, 245] using a variety of features including word frequencies, sentence structure or dependency trees. Adding these approaches to *OntoSuite-Miner* would allow the capture of an extra layer of data of the gene association and potentially increase the usefulness of the resulting gene annotation data.

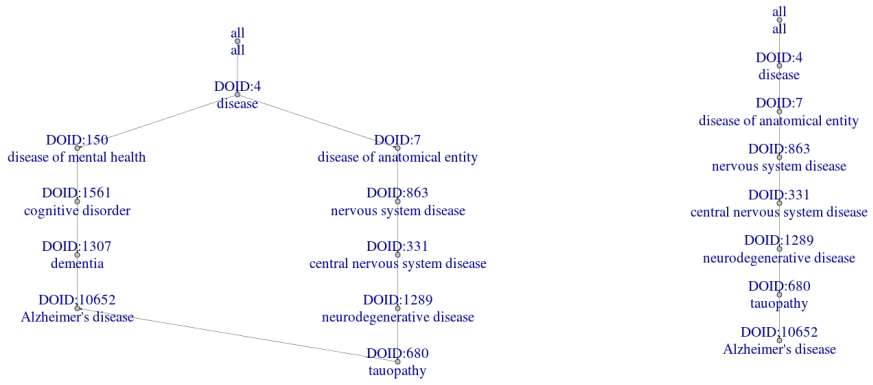
There are other aspects of the implementation that could be improved. In section 2.2.2, genes were linked to diseases by SNP based on their location on the chromosome. The closest up/down stream and the overlapping genes of a SNP were considered important for the disease/phenotype. This sequence level prediction assumes that sets of alleles on the same small chromosomal segment tend to be transmitted as a block through generations thus SNPs are likely to be inherited together with their closely located up-stream/down-stream genes during evolution. The current implementation uses the closest stream/down-stream genes regardless their distance to the SNP. This could be improved by adding a score to such link, taking into account the distance. On the other hand, instead of using sequence level information, transcript level or protein level prediction could also be integrated to improve the quality of the data but is not currently implemented. It would be interesting to incorporate such information and there are many tools available for this task. Variant Effect Predictor(VEP) [246] uses a rule-based approach to predict the effects that each allele of the variant may have on the transcript using a set of consequence terms from the Sequence

Ontology(SO) [247]. SnpEff [248] predicts the effects of variants on genes. SIFT [249] predicts whether an amino acid substitution is likely to affect protein function based on sequence homology and the physico-chemical similarity between the alternate amino acids. Other protein level predictions tools including PolyPhen [250] (predicts the SNP effect on the structure and function of a protein) and PANTHER Coding SNP Analysis tool [251] (estimates the likelihood of a SNP causing a functional impact on the protein) are freely available to use.

Lastly, the ontology used by *OntoSuite-Miner* can affect the performance and as a result, affect the resulting annotation dataset. Thus, there is always room for improvement and refinement of the ontology itself. In this project, the Human disease Ontology was used to represent disease concepts. HDO was developed based on a subset of the UMLS disease concepts. It is a relatively new ontology (published in [39] in 2012) aiming to provide an open source ontology for the integration of biomedical data that is associated with human disease. Thus, it is frequently updated and refined in both terms and the ontology structure itself. New terms may be added while existing terms may be refined to better represent the disease concept. For example, the disease term 'DOID:10652 Alzheimer's disease' was a child term (more specific term) of 'DOID:680 tauopathy' and 'DOID: dementia' but was removed from the later after HDO version 1.2 release on 2015-12-04 (Firgure fig. 2.30 on the facing page); 'pericentral pigmentary retinopathy' was added to term ' DOID:10584 retinitis pigmentosa' as an exact synonym. Moreover, some of the errors identified by *HDGDB* in this thesis suggest that additional research is needed into the definition of the ontology to eliminate ambiguities. For example, 'OCD' for DOID:84 osteochondritis dissecans(OCD) and 'obsessive-compulsive disorder' or 'AD' for DOID:10652 Alzheimers disease(AD) and autosomal dominant. These changes in the ontology not only have an impact on OntoSuite-Miner, but also affect some of the algorithms in OntoSuite-Analytics (discussed in the next chapter) where ontology structure is taking into account, for instance, when using the elimination algorithm in enrichment analysis. Therefore, a corresponding update needs to be scheduled when either the data sources or the ontology have a major new release.

Other potential usage of *HDGDB*

Ontology based gene annotation enables a wide rage of modern bioinformatics analysis. *HDGDB* potentially could be used in other disease network analysis which



(a) (b)

Figure 2.30: 'Alzheimer's disease' in HDO ontology (a) before and (b) after version 1.2 release on 2015-12-04. HDO is frequently updated and refined, and changes to the ontology structure were made. New terms may be added while existing terms may be refined to better represent disease concepts.

may provide insights into the interconnections between human diseases. We are currently witnessing a shift from a 'single gene single disease' paradigm towards an 'interplay of different disease modules' [204, 205]. It is difficult to consider diseases as being invariably independent of one another. In fact, disease modules can overlap, so that the underlying molecular mechanism causing one disease can affect other disease modules. Disease network analysis has emerged as a powerful way of studying the inter-relationship between diseases and their associated genes. Such networks of diseases and disease genes offers a platform to explore all known diseases and disease gene associations in a single graph theoretic framework, and has already been used to reveal the common genetic origin of many diseases as well as predict potential disease gene candidates [204, 206, 207].

In a typical disease network, nodes represent diseases and links represent various molecular relationships between them. Some commonly used relationships which are based on shared features between diseases include their shared causal genes, their shared regulatory microrRNAs, related pathways and various other external influences. Details of some of these approaches have been discussed by Barabasi et al. in [204].

Shared Gene approach A gene associated with two different diseases is often an indication that the two diseases have a common genetic origin. Disease nodes in the network are linked if they share one or several genes. Goh et al. [206] used data from the OMIM database to build a network of disease containing 1284 diseases, out of which 867 were linked to one or more other diseases. 516 diseases were found in a single disease cluster, forming the largest connected component in the disease network. It was observed that similar diseases from the same class (indicated by color in the figure) were more likely to share genes than diseases that belong to a different class. For example, cancers form a tightly interconnected sub-cluster by a small group of genes associated with multiple cancers including *P53*,*KRAS*,*ERBB2* or *NF1*. Indeed, Park et al. [252] showed co-morbidity between linked diseases from the observation that patients with a primary disease are twice as likely to develop a secondary disease if the secondary disease shares genes with the primary one. However, such compatibility is not guaranteed between linked diseases. This was partly attributed to different contextual scenarios of their genetic mutations where mutations on the same gene can have different effects on the function of the gene product or on its organ-based expression.

On the other hand, a gene network can be constructed with a similar approach where nodes in the network represent genes and they are linked if they associate with the same disease(s). Such networks have been explored and have resulted in a number of different finding regarding predicting biological functions, for example, in recent studies by Gillis and Pavlidis in [253–255].

Shared metabolic pathway approach An enzymatic defect that affect a metabolic reaction may potentially affects all downstream metabolic reactions in the same pathway, leading to metabolically-induced disease phenotypes. This is particularly true for metabolic diseases. In this scenario, nodes in the network represent diseases and two diseases are linked if the enzymes associated with them catalyze adjacent reactions. Such a network has been used to study disease co-morbidity. For example, Lee and Chung [256] constructed a metabolic disease network(MDN) and showed a 1.8-fold co-morbidity increase in diseases linked in this network when compared to those that are not.

Shared microRNAs approach A microRNA (abbreviated miRNA) is a small non-coding RNA molecule that functions in RNA silencing and post-transcriptional regulation of gene expression. A signal miRNA down-regulates potentially hundreds of

target mRNAs, and often play a key role in cellular functions such as development, differentiation, proliferation, apotheosis and metabolism. Recently, Lu et al. [257] implemented miRNA regulation into disease networks, where disease nodes were linked when their associated genes are regulated by one or more shared miRNAs. The resulting network showed segregated disease clusters at the miRNA level for cancer and cardiovascular diseases.

Such 'guilt by association' approaches link two diseases in the network by some type of shared feature between them. The shared gene approach is an excellent approach to explore the inter-relation between human diseases and the corresponding disease genes for gene disease association data like *HDGDB*. A human disease network (HDN) and disease gene network (DGN) can be constructed by building a bipartite graph linking diseases to genes using HDGDB. The bipartite graph contains two sets of nodes, disease nodes and gene nodes, such that every edge in the graph connects a disease node to a gene node. From this bipartite graph, two projections can be made: 1) HDN, where nodes are disorders and a pair of disorders is linked if they share at least one associated gene is known to be involved in both disorders and 2) DGN, where nodes are genes and a pair of genes is linked if they are both involved in at least one shared disorder. The resulting HDN and GDN can be used to explore disease-disease or gene-gene interconnections in the same way as proposed by Goh et al. [206].

Chapter 3

An R package for generalized ontology term enrichment analysis

3.1 Background

The development of high-throughput genomic, proteomic and bioinformatics scanning approaches allow researchers to simultaneously measure the properties of genome-wide genes and proteins across entire genomes. Large interesting gene lists are often generated from these high-throughput approaches, but the biological interpretation of these potentially interesting genes is still a challenging and daunting task. The gene-annotation enrichment analysis is a promising strategy for functional analysis of such gene lists, making it possible to systematically dissect them to gain biological insight, and has became a standard practice in the downstream analysis of high-throughput approaches.

Several bioinformatics enrichment tools have been developed during the last decade. In a survey, Huang et. al. [16] identified and reviewed at least 68 different enrichment methods. Despite the distinct features of the different enrichment tools, the general procedure for the enrichment analysis is similar and can be summarized into three parts:

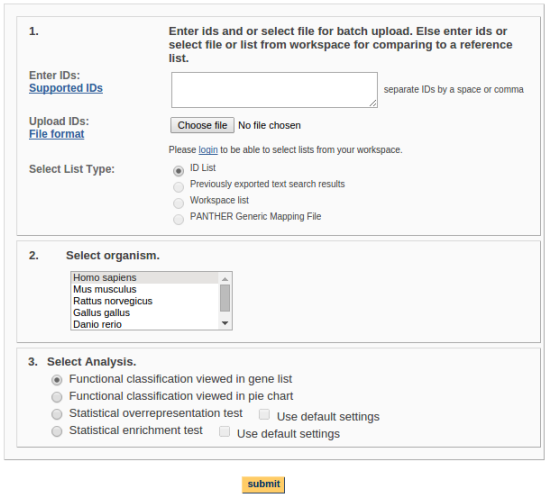
1. Preparation of the backend annotation database. This usually refers to ontology based annotation where genes are annotated with predefined ontology terms. For example, the Gene Ontology [20] represents over 40,000 biological concepts, describing how genes encode biological functions at the molecular, cellular and tissue system levels. The GO annotation links genes with GO terms indicating the various functions that a particular gene may have.

2. Calculation of enrichment (algorithm and statistics) against a reference set; For example, if 20% of the genes under study are found to be annotated with "*synapse complexity*" compared to 8% of the genes in the human genome, this enrichment can therefore be assessed by some kind of statistical methods. The pre-eminent statistical test is the Fisher's exact test from which an enrichment p-value is calculated indicating the statistic power of the test. Thus a conclusion may be drawn that "*synapse complexity*" is an enriched annotation and therefore may play an important role in the system.

3. Post-process and presentation of results.

The Gene Ontology has been, and is still the most, if not the only ontology used in enrichment analysis largely due to the limitation of annotation coverage discussed earlier. The GO Consortium's AmiGO [258] is provided as a web application that allows functional enrichment analysis for user uploaded gene lists. DAVID [53] provides a comprehensive set of functional annotation tools the interpretation of biological meaning behind list of genes, including identifying enriched biological themes, particularly GO terms A set of related bioinformatics tools such as the Gene ID Conversion Tool is also freely available as a web application. These tools provide easy to use enrichment analysis for users but are limited for advanced usage scenarios such as batch analysis or when parameters need to be customized. Various programming packages support enrichment analysis, such as topGO [33], GOstats [90], CompGO [259] for GO terms, GAGE [72] and Pathview [260] for pathways and DOSE [261] for the diseases. These tools play an important and successful role in functional analysis of gene lists for various biological studies [49, 53–59].

Despite the usefulness of existing enrichment analysis tools and the different statistical algorithms used for finding enriched ontology terms, they mostly only work for a specific type of ontology, GO in the majority of cases, even though the general underlying enrichment analysis process is in principle similar for any ontology. A standard framework for general ontology enrichment analysis is needed. In the following sections, I proposed *topOnto*, an analytic R package, integrating a range of statistical algorithms and topology methods for ontology enrichment analysis that aims to ameliorate the current limitations of enrichment tools, and facilitate generalized ontology enrichment analysis.



(a)

Term	Background frequency	Sample frequency	Expected	+/-	P-value
synaptic transmission (GO:0007268)	588	44	4.369e+00	+	1.241e-28
establishment of localization (GO:0051234)	3552	86	2.639e+01	+	2.391e-24
transport (GO:0006810)	3469	85	2.577e+01	+	2.606e-24
cell-cell signaling (GO:0007267)	914	47	6.791e+00	+	7.656e-24
single-organism cellular process (GO:0044763)	10724	143	7.968e+01	+	2.150e-23
single-organism localization (GO:1902578)	2972	76	2.208e+01	+	5.947e-22
single-organism transport (GO:0044765)	2796	73	2.077e+01	+	2.489e-21
single-organism process (GO:0044699)	12106	148	8.995e+01	+	3.637e-21
localization (GO:0051179)	4323	89	3.212e+01	+	1.898e-20
ion transport (GO:0006811)	1177	45	8.745e+00	+	1.326e-17

(b)

Figure 3.1: (a) The GO term enrichment analysis provided by Amigo and (b) the enrichment result of a toy example gene list.

3.1.1 Organization of the chapter

In this chapter, in order to adjust the second part of the TBI challenge which is the availability of the appropriate bioinformatics tools for the analysis and exploitation of human disease data, I proposed the second part of the *OntoSuite* framework, the *OntoSuite-Analytics*, which consists of a set of R packages including *topOnto* and the corresponding data packages for the ontologies. I provided implementation details, the main design decisions and a number of validation approaches for the package. I discussed statistical limitations, back-end annotation database, multiple hypothesis correction and the key design decisions to ameliorate them. A modified GSEA algorithm, named GSEA-CSW was proposed for gene set enrichment analysis. The algorithm was implemented and tested with syntactic gene expression data. A practical use-case example of the tool was presented with the investigation the activity-regulated cytoskeleton-associated protein (ARC) complex followed by a conclusion section.

3.2 Implementation of *topOnto*

topGO is a powerful gene list enrichment analysis package, available from R/Bioconductor, designed to facilitate semi-automated enrichment analysis for Gene Ontology (GO). It provides unified enrichment analysis framework that facilitates comparison between different enrichment methodologies. In addition, it implements a set of topo-

logical algorithms proposed by Alexa et al. [33] which calculate enrichment for GO terms while accounting for the topology of the GO graph. The package has been widely used and plays an important and successful role in various high-throughput biological studies [262–265]. However, *topGO* only supports Gene Ontology analysis. As discussed before, the ability of using different ontologies would be beneficial for biological interpretation. Cross comparison of results across several ontologies can provide further evidence of important biological processes if those processes are represented in different ontologies. Fortunately, ontologies are usually defined in a similar structure and stored in text files following the same standard OBO format. This motivated me to adapt *topGO* to *topOnto*, extending of all the functionalities of the origin *topGO* to any ontology of interest. Because *topOnto* was implemented on top of *topGO*, there is a large amount of code being reused and adapted from *topGO*. The detailed implementation of the *topGO* has already been described by its author Alexa in [266], thus not repeatedly described in this thesis. I will only focus on the implementation and usage of *topOnto* that differs from the *topGO*. *topOnto* is freely available on github (https://github.com/statbio/topOnto). A high-level schematic describing my implementation of the *topOnto* package is shown in fig. 3.2. The work flow of *topOnto* can be divided into four steps: i) Ontology preparation; ii) Data preparation; iii) Running the enrichment tests and iv) Analysis/comparison of the results.

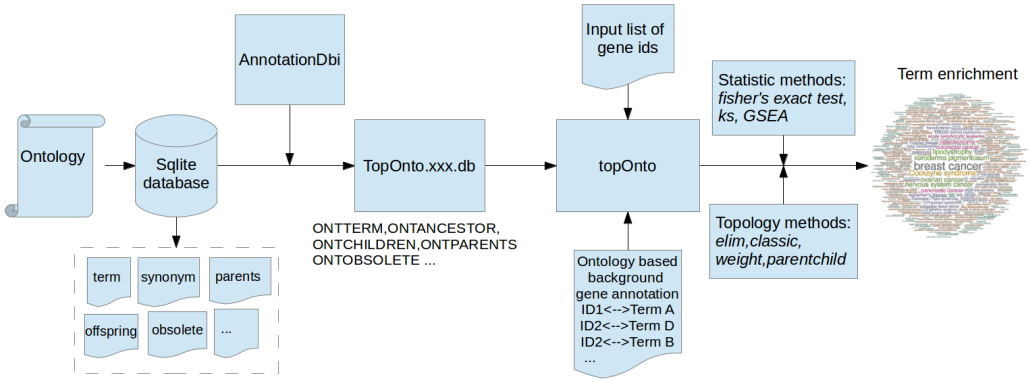


Figure 3.2: The work flow of *topOnto* can be divided into four steps: i) Ontology preparation; ii) Data preparation; iii) Running the enrichment tests and iv) Analysis/comparison of the results.

3.2.1 Ontology preparation

To use *topOnto* for enrichment analysis, firstly one needs to decide what ontology to use and prepare the corresponding ontology objects. The ontology is usually defined in an OBO flat file. Scripts are provided by *topOnto* to extract information from the OBO file and store it in a SQLite database following the *AnnotationDbi* bimaps database schema [267]. For example, the SQLite table *term* stores the ontology term id, name and definition while the tables *parents* and *offspring* store the interrelationships between terms. The ontology package implements the *AnnotationDbi* interface, which serves as an abstraction layer above the underlying SQLite database, providing methods to query the database and generate objects that allow easy access from within R. Objects such as *ONTTERM, ONTANCESTOR* and *ONTCHILDREN* are provided to represent the ontology hierarchy. The ontology package is named *topOnto.X.db*, where *X* stands for the name of the ontology, indicating that this is an ontology package providing information needed by the *topOnto* package. For example, *topOnto.HDO.db* for the Human disease ontology.

The advantage of using such an abstraction layer is that, regardless of the different ontologies, the same set of objects will be generated representing the corresponding ontology. This way, *topOnto* doesn't need to know the details of the underlying SQLite database and can work generically with a set of objects provided by the *topOnto.X.db* package for each ontology. This is one of the most important features of *topOnto* which making it possible to apply the same set of enrichment methods to different ontologies.

3.2.2 Data preparation

An R object of class *topONTdata* was defined in *topOnto*, which is designed to be a master object that stores the ontology, gene annotations and test statistic. To construct this object, one first needs to define the list of genes of interest. This can be a preselected list of gene ids or gene expression data with a criteria for selecting genes based on their scores. Secondly, a gene universe is to be defined which serve as a reference background when calculating the terms enrichment statistics. See section 1.2 for how to define a proper gene universe. Next, the ontology is loaded into the *topONTdata* from *topOnto.X.db* packages where the ontology is internally represented as a directed acyclic graph (DAG) (fig. 3.3). In the ontology DAG, individual terms are represented as nodes connected, by directed edges, to more specific nodes, such that each node is a more specific child of one or more parents. Note that for such a graph, the notion

of child and parent can be used, where a child term is more specific than its parent. Directed edges between nodes represent their relationship. The ontology based gene annotation can be loaded into the DAG from a tab delimited file in which each row represents an ontology term and its annotated genes. *topOnto* currently provides annotation for Human Disease Ontology (HDO), Human Phenotype Ontology (HPO), Gene Ontology (GOBP, GOCC, GOMF)[1], the Reactome Pathway Ontology (RPO)[2], Panther protein class Ontology (PCO)[3] and the Chromosome Ontology (CO)[4]. The HDO and HPO annotations were generated by *OntoSuite-Miner* while the others were taken from their website. Note that strictly speaking, the RPO and CO is not a formally defined ontology but a list of controlled vocabularies structured hierarchically. It has the basic structure of an ontology (terms, relationship between terms) which makes it possible to apply the same methodologies as for a standard ontology.

In the following sessions, I will be using a small part of the Human Disease Ontology as an example to illustrate the different methods and algorithms implemented in *topOnto*. There are 21 nodes and 20 links in this sub HDO DAG. Each node represents a HDO term and each the link represents a 'is_a' relationship. The directed edges are pointing from parent nodes to child node. The DAG has 8 levels and 5 leaf nodes. The root node is 'all'. Its structure is shown in fig. 3.3.

[1] http://geneontology.org/page/download-annotations
[2] http://www.reactome.org/pages/download-data/
[3] http://www.pantherdb.org/
[4] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/

Figure 3.3: DAG structure showing part of the Human Disease Ontology. There are 21 nodes and 20 links in this sub HDO DAG. Each node represents a HDO term and each the link represents an 'is_a' relationship. The directed edges point from parent node to child node. The DAG has 8 levels and 5 leaf nodes. The root node is 'all'.

Gene are assigned to the nodes in the DAG according to the gene annotation provided, followed by a 'roll-up' procedure in which a gene assigned to a node, will also be assigned to all its direct parent nodes. This result expands the gene annotation base on the hierarchical structure of the ontology DAG, aggregating genes from the bottom (specific nodes) to the top (general nodes) of the structure. The 'roll-up' process is based on the True Path Rule of the ontology definition where an annotation for a term(node) in the ontology hierarchy is transferable to its ancestors (transitivity). Thus, if a gene is annotated to a term(node), it is also implicitly associated with all the terms on its True Rule (see section 1.1.1.4 on page 6) [268]. In the Gene Ontology biological process, for example, if a gene is involved in 'developmental programmed cell death', then it is reasonable to say that it is also involved in the parents process such as 'cellular developmental process' or 'cell death'. Due to the unbalanced nature of research interests, some areas are better studied than others. In addition, genes are sometimes annotated with different granularities; a gene associated with Parkinson's disease may be annotated with a specific type of Parkinson's disease, for example autosomal dominant Parkinson disease 1, or with a more general disease category like neurodegenerative disease or nervous system disease. This results in an unbalanced

ontology annotation where some of the less studied terms are poorly annotated, thus unlikely to be identified as enriched in the analysis. The 'roll-up' process was designed to avoid this problem by aggregating the effect of poorly annotated terms to their parent terms so that these joint effects are more likely to be identified. However, the True Path Rule introduces dependency into the ontology terms which was referred to as the 'inheritance problem' by Grossmann et al. in [269] , which may result in some of the statistical methods producing misleading results.

By default, all terms available in the ontology are used to construct ontology DAG. Optionally, *topOnto* allows the user to define a domain specific subset of ontology terms (a 'clip') before constructing the ontology DAG so that all the following statistical analyses are performed with this subset rather than the full ontology. The advantage of using a more specific pruned ontology in enrichment analysis was discussed in section 1.1.1.5.

A single R object of class *topONTdata* contains the gene data (genes of interest/-gene universe), ontology data and gene annotations. Arguments are available for controlling the construction of the *topONTdata* object. For example, the argument *ontology* specifies which ontologies to be used while the argument *nodeSize* is used to prune the ontology hierarchy of terms which have only a few, for example, less than 10, annotated genes. A summary of the *topONTdata* object can be seen by typing the object name at the R prompt. Having all the data stored in this single object facilitates easy access to identifiers, annotations and to basic data statistics.

3.2.3 Running the enrichment tests

Once a *topONTdata* object has been created, it is possible to process it with a number of different statistical algorithms and topology methods. The statistical algorithms are used to measure the significance (p-value) of the enrichment while the topology methods aim to reduce the number of false positive results by rating into account the ontology structure. *topOnto* currently supports statistical algorithms including Fisher's exact test, KolmogorovSmirnov test and GSEA. Topology methods including 'classic', 'elim', 'weight01' and 'parentchild' are also available and are described further below. table 3.1 presents the compatibility between the statistics and the topology methods.

A function *runTest* is implemented to apply the specified test statistic and topology method to the *topONTdata* object. It returns an object of class *topONTresult* for each test statistic/topology method pair. The results might be different when using different

	classic	*elim*	weight	weight01	parentchild
fisher	✓	✓	✓	✓	✓
ks	✓	✓	✓	✓	–
GSEA	✓	✓	–	–	–

Table 3.1: Algorithms/topology methods currently supported by *topOnto*.

statistic/topology combinations. A high-level interface script *run.batch* is used to apply the specified test statistics and methods to the data simultaneously across multiple ontologies. This returns a list of results indexed by ontology names. It is up to the user to decide which combinations are suitable for their data. A brief introduction is given which can be used as a guide to choose statistic and topology methods and to aid interpretation of the results.

3.2.4 Statistical algorithms

The Fishers exact test is a count based statistical significance test used in the analysis of contingency tables. It is useful when only a list of interesting genes is provided but no further information is available (gene scores, weighted, gene expression measurements etc.). Typically this gene list is preselected from other studies such as genes that belong to a certain cluster or top ranked genes in a microarray experiment that are over expressed in one group against another. In the Fisher's exact test, every gene contributes equally. Sometimes genes expression measurements or other properties like score or weight of the genes are available. For example, a sore can be calculated for each gene based on the expression value in a microarray experiment. The score indicates how distinguish/important the gene is within the two different test groups. In this case, score based algorithms like the KolmogorovSmirnov test or GSEA are more appropriate because these algorithms take all the gene scores into account, allowing the more highly score genes to contribute more to the final enrichment result.

3.2.5 Topology methods

The *classic*,*elim*,*weight* and *weight01* methods were introduced by Alexa et al. in [33]. The *parentchild* method is proposed by Grossmann et al. in [269].

The *classic* method, differs from the other methods, as it does not take the ontology structure into account, treating all the ontology nodes as a flat list and applying the

chosen statistical algorithm independently to each node. Another way to think of the *classic* method is that when testing for a particular node, it is actually abstracting the DAG to the level of that node with all the information below that level aggregated to it. As a result, the top enriched nodes are usually often generic terms due to the 'roll-up' process. These terms reflect the joint effect of all their offspring terms and give a top-down overview of the genes in relation to the ontology. In practice, the *classic* method identifies the upper bound of the significant value of a term when there is no gene annotated to any of its children terms. As the number of children annotated grows, the significant value is often over-estimated because the annotations aggregated are evaluated multiple times(see fig. 3.6 for an example).

```
markedGenes ← Ø:   nodeSig ← Ø
get the DAG levels list   dagLevels
for i from max(dagLevels) to 1
  for u in nodes(dagLevels, i)
    genes[u] ← genes[u]\markedGenes[u]
    nodeSig[u] ← FisherTest(genes[u], sigGenes)
    if nodeSig[u] ≤ thresheold then
      for x in upperInducedGraph(u)
        markedGenes[x] ← markedGenes[x] ∪ genes[u]
  end
end
return   nodeSig
```

Figure 3.4: The gene elimination algorithm [33]. Genes are removed from the node and all its ancestors nodes in the ontology DAG when the nodes is marked significant. Thus the algorithm always detects the most specific terms in the ontology hierarchy.

The *elim* method processes the ontology terms by traversing the ontology hierarchy from bottom to top, i.e. it first assesses the most specific (bottom-most) nodes, and proceeds later to more general (higher) nodes. If a node is found significantly enriched, all the gene annotated to this node will be removed from all its ancestor nodes. To be precise, as shown in fig. 3.4, the algorithm keeps a history of the significant nodes and their genes in *markedGenes*. Every time a node u is being tested, a gene elimination process takes place before the significance test is performed, pruning the genes annotated to u, denoted $genes[u]$. If node u is found significant then all of the genes mapped to it are marked as removed in all the ancestor to be node u. The algorithm recursively walks up the ontology structure until all nodes have been processed (fig. 3.5). The

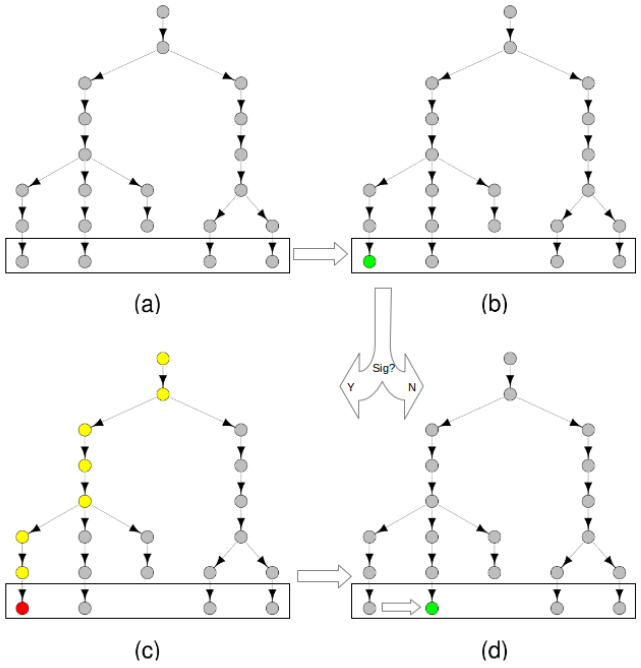
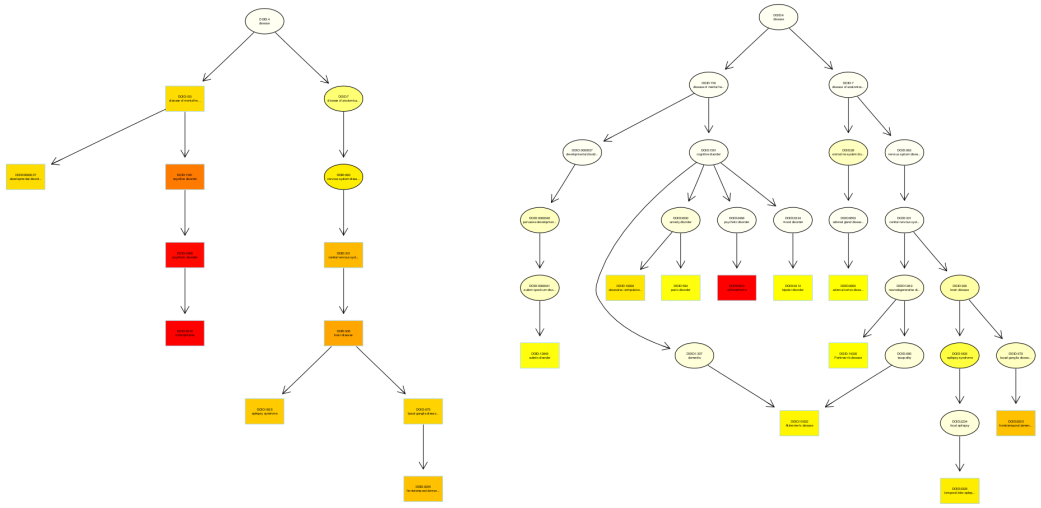


Figure 3.5: The *elim* method. (a) The process starts from the bottom nodes of the ontology DAG. Since nodes from the same level share no edge, they can be investigated independently. (b) For a particular node (green), genes that have been marked as removed in a previous step are removed. A chosen statistical method is applied to calculate the enrichment significance level, i.e. the p-value. (c) Node is marked significant (red) if its p-value is smaller than a previously defined threshold. Genes annotated to this node are marked to be eliminated in all of the ancestors (yellow) up to the root. (d) The process moves to the next node.

elimination process effectively prevents genes from repeatedly contributing to the significance of the nodes across the ontology hierarchy, thus, in another words, it always tries to find the most specific enriched nodes. As a result, the selected enriched nodes usually provide a bottom-up perspective regarding to the ontology (fig. 3.6). The *elim* method minimizes the number of false positives. However, it does tend to miss some true positives at higher (more general) levels of the ontology hierarchy. Another limitation of the elim process is that it, on top of the 'roll-up' process, induces another layer of dependence between the ontology terms, which may affect subsequence analysis such as applying multiple hypothesis testing correction on the results. The elim method is the default method implemented in *topOnto*.



(a) (b)

Figure 3.6: The comparison of enrichment result (top 10 enriched terms from the a toy example list of genes) distribution of (a) the *classic* method and (b) the *elim* method in the HDO ontology DAG. The color indicates the significance of the nodes. Colored nodes are found significantly enriched with a $p - value <= 0.5$, while red nodes are more significant (smaller p-value) then yellow nodes. If a term contains the same genes as one of its children (due to the 'roll-up' process), the classic method always gives the same score to both terms while the child term in this case often contains more specific biological information, thus is likely to be more interest.. The *elim* method is designed to tackle this problem, by computing the significance of a term dependent on the significance of its children. It effectively proves the same gene contributes to the enrichment result multiple times along the path to the ontology root. As a result, the classic method usually generates enriched terms that are located narrowly and vertically in the ontology DAG, while the *elim method* tends to identify enriched terms that are horizontally spread in the ontology DAG(see table 3.5 for an example).

Instead of completely removing the genes that are annotated with significantly enriched descendant terms, the more involved *weight* method uses a similar bottom-up strategy as the *elim* method but assigns weights to genes as a function of the scores of neighboring terms. The algorithm is shown in fig. 3.7. In the *weight* method, the ontology terms are processed bottom-up level by level as in the *elim* method. Initially, all weights for the genes annotated to a node are set to 1. For a particular term *u*, a modified version of the Fisher's exact test is used to determine the term's significant value. If the term is found more significant than its children terms, genes in the children are down-weighted and significant scores are recalculated base on the new gene weight, resulting a decreased significance of the children. On the other hand, if any of the children are more significant than the parents *u*, the genes common to the child and *u* are down-weighted in *u* and its ancestors, decreasing the significance of term *u*. The principle behind the *weight* method is to reinforce differences in significance between u and its neighbours, thus enhancing the ability of detecting the most significant local terms in the ontology hierarchy. It is less strict than the *elim* method, and tends to miss less true positives [33]. The *weight01* method is a mixture of the *elim* and *weight* methods. Here, the gene elimination process only happens when a term is not the most significant local term.

The *parentchild* approach is developed specifically to avoid the so-call 'inheritance problem' of Grossmann et al. in [269]. The rationale behind this method is that, due to the 'true path rule', when multiple terms are tested simultaneously, the chance of a term being enriched is much higher if one or more of its parental terms is also enriched. To avoid this problem, the significance level of a given term is computed by taking into account the immediately more general terms (the parents). This can often produce less redundant result and lead to the removal of false positives, since some of the more specific terms are eliminated if their parent is determined to be significant. Note the difference between the 'bottom-up' methods and *parentchild* methods is that the former tend to eliminate parent terms while the later tend to keep them.

Simulation results reported by Alexa et al. [33] show that the *weight* algorithm has less false positives than the *classic* method and misses fewer true positives, while *elim* has even less false positives than *weight* but misses more true positives. Grossmann et al. in [269] compare the *elim*, *weight* and *parentchild* algorithms with simulated GO data and conclude that each method has its own advantages in certain scenarios. The advantage of *topOnto* is that it provides a platform, allowing simultaneously testing of enrichment with different statistic/topology methods. Users can easily add new

```
for u in nodes(dag)   nodeW[u] ← 1   end
nodeSig ← Ø
get the DAG levels list   dagLevels
for i from max(dagLevels) to 1
  for node in nodes(dagLevels, i)
    computeTermSig(node, children(node))
end
return   nodeSig

function computeTermSig(u, children)
   nodeSig[u] ← WFisherTest(genes[u], node W[u])
   if children = Ø then return fi
   for ch in children
     weights[ch] ← sigRatio(nodeSig[ch], nodeSig[u])
   end
   sigChildren ← {ch | weights[ch] ≥ 1, ch ϵ children}
   if sigChildren = ∅ then                          /*CASE1*/
     for ch in children
       node W[ch] ← nodeW[ch] ⊗ weights[ch]
       nodeSig[ch] ← WFisherTest(genes[ch], node W[ch])
     end
     return
   fi                                               /*CASE2*/
   for ch in sigChildren
     for w in upperInducedGraph(u)
       nodeW[w] ← nodeW[w] ⊗ 1/weights[ch]
   end
   computeTermSig(u, children \ sigChildren)
```

Figure 3.7: The weight algorithm [33]. Weights are assigned to genes as a function of the scores of neighboring terms. If the term is found more significant than its children terms, genes in the children are down-weighted and significant scores are recalculated base on the new gene weight, resulting a decreased significance of the children. The same applied when any of the children are more significant than the parents, which results in detecting the most significant local terms in the ontology hierarchy.

methods into the platform and test them against the other methods. It is recommended to test the same set of genes with multiple methods and consider differences in the results when making any conclusions.

3.2.6 Analysis of the results

An object of class *topONTresult* is returned for each statistic/topology method pair containing the p-values for each node in the ontology DAG. The p-values returned are raw p-value without any multiple testing correction. This is because 1) In many cases a FDR/FWER adjustment procedure can produce very conservative p-values and declare no, or very few, significant terms. This can lead to an increase of false negatives, the loss of interesting terms that contain valuable information; and 2) the gene 'roll-up' process can induce dependence into the terms, thus the independent assumption of multiple testing does not directly apply. Some of the topology methods including *elim,weight* and *parentchild* compute the term significance conditioned on the neighbor terms. However, the user can perform an adjustment on top of the of *topONTresult* object if it is considered important for the analysis. The raw p-value was provided in the *topONTresult*, together with the corrected p-value using the default method of the Benjamini-Yekutieli(2001) [119] FDR correction.

A Function *GenTable* is implemented to merge the results into a summary table where p-values for each method are put together for easy comparison (table 3.2.). Certain patterns can be found when comparing p-values across different topology methods (see fig. 3.12a on page 145 for some examples). The significance of genes aggregate from bottom to top due to the gene 'roll-up' process resulting in a vertical decrease of significance from top to bottom of the DAG. For the *classic* method, this leads to identifying significant terms located in the same vertical branch of the ontology DAG. For to the same reason, the *parentchild* method usually identifies a smaller number of low level terms that horizontally spread across the ontology DAG. The elimination methods including *elim* and *weight01* remove significant genes from bottom to top which usually results in a vertical increase of significance from top to bottom of the DAG. These methods tend to identify high level enriched terms that horizontally spread across the ontology DAG. There are circumstances when the *classic* approach and the elimination approach generate the same p-value for a term. This indicates that the topology structure has no effect on this term, either because this term is a leaf term (has no descendants) or none of its descendants are significant. When the *classic* ap-

proach identifies a significant term but the elimination approach does not, it suggests that this term is a relatively low level term (has many descendants) and its significance does not come from the genes annotated directly to it but the genes aggregated from its descendants.

	TERM.ID	Term	Level	Annotated	Significant	classic	elim	weight	parentchild
1	DOID:5419	schizophrenia	6	1848	48	9.6e-22	9.6e-22	9.6e-22	0.76922
2	DOID:2468	psychotic disorder	5	1858	48	1.2e-21	1.0000	1.00	5.3e-06
3	DOID:1561	cognitive disorder	4	3037	54	2.8e-17	0.6510	1.00	0.00087
4	DOID:936	brain disease	6	2001	44	6.8e-17	0.1099	1.00	9.9e-07
5	DOID:150	disease of mental health	3	4787	66	1.8e-16	0.2115	0.69	1.8e-16
6	DOID:1826	epilepsy syndrome	7	685	25	3.8e-14	0.0011	6.7e-08	0.00159
7	DOID:331	central nervous system disease	5	4834	61	8.0e-13	0.4453	1.00	0.00240
8	DOID:863	nervous system disease	4	6684	70	2.5e-11	0.0991	0.94	3.5e-07
9	DOID:1443	cerebral degeneration	7	282	14	5.4e-10	5.4e-10	3.8e-09	0.00190
10	DOID:9255	frontotemporal dementia	8	197	12	1.0e-09	1.0e-09	1.0e-09	0.24237

Table 3.2: A summary table generated by the function *GenTable* of *topOnto*. The first row of the result can be interpreted as follows: HDO term 'DOID:5419 schizophrenia' is annotated to 1848 genes among which 48 appear in the given list of 'interesting genes' as significant gene. p-value is calculated by Fisher exact text with four different topology methods, namely *classic,elim,weight* and *parentchild* and present together for easy comparison.

3.2.7 Weighted GSEA

In chapter 2 on page 33, a data mining framework is built to generate a human disease ontology (HDO) gene annotation from publicly available databases,including GeneRIF, OMIM and Ensembl Variation. The resulting gene-HDO annotation is scored base on a confidence level. A score of 1 means that an annotation is only found once, in one of the three sources by either MetaMap or NCBO-Annotator, indicating that this annotation is not reliable and should be used with caution. A higher score means more evidence exists and the association trustworthy. This could for example mean that it has been found in multiple corpora, multiple source, by both MetaMap and NCBO-Annotator. Details of the annotation confidence score is described in section 2.3.2.

As discussed previously, the quality of the gene annotation has a great effect on the final enrichment result. However, none of the currently available gene annotation databases are complete or error free. There is always a trade off between annotation coverage and accuracy. Manual annotations are considered accurate most of the time.

However, because of the time lag necessary for the manual curation process, recent annotations are missing and the overall coverage are usually low [44]. On the contrary, annotations generated with automatic methods (without human expert involvement) may have quality issues but usually have good coverage. For example, in the Gene Ontology, out of 481685 total annotations available for *Homo sapiens*, 155499 (32%) are inferred exclusively from electronic annotations (with Evidence Codes *IEA*, no human expert involvement for checking the annotation's accuracy). Even though some of these *IEA* annotations are incorrect [84,85], the vast majority of them are reasonably accurate [83]. However, to the best of my knowledge, at the present time, none of the enrichment methods allow any type of weighting by annotation confidence which is a limitation since manual curations are generally more trustworthy. In order to solve this issue, and to make better use of the annotation data discuss in section 2.3.2 on page 66, for the first time, I propose a modification of the original GSEA methods referred GSEA-CSW (confidence score weighted), implemented in *topOnto*, allowing the use of weighting based on the annotation confidence, which as a result, increases the magnitude of the effect of better annotations and decreases the magnitude of the effect of weaker ones.

The biological utility of GSEA can be improved by including additional biological features. Utilizing more domain knowledge is likely to reveal more insights from the analysis. In [270], Jun et.al integrate KEGG PATHWAY information to weight genes involved in pathways. An appearance frequency (AF) is assigned to each gene based on the number of times it appears in all of the KEGG pathways. Those genes that are involved in many pathways are usually responsible for housekeeping functions, thus receive a high AF whereas other genes which are more specialized and play unique roles in one or a few pathways, receive a low AF. GSEA with the integration AF scores is claimed to perform better both statistically and biologically [270]. Several other recently introduced techniques incorporate concepts of gene topology, including ScorePAGE [271], gene network enrichment analysis [272] and network topology analysis [273]. These additional features improve the GSEA methodology not just conceptually but practically improve the enrichment results. Similar to the above approaches, GSEA-CSW incorporates the quality of the annotation into the enrichment analysis with the rationale that high scored genes should be contributing more to the enrichment analysis.

Recall that in GSEA, an enrichment score (*ES*) is calculated by walking down a ranked gene list L, increasing a running-sum statistic when encountering a member

gene of set S (P_{hit}) and decreasing it when encountering a non-member gene of set S (P_{miss}) (Equation eq. (1.1) on page 18). The magnitude of the increment during the running-sum process is controlled by an exponent p. Typically, all genes in a gene set S are treated equally in the analysis. However, in the case of using automatically generated annotations, genes with a lower confidence score (CS) may not be as reliable, thus should have less effect than those have a higher confidence score. For example, for a gene set S, if most of the top ranked genes are low confident annotated gene, then set S might be enriched but with a high chance of being a false positive result. GSEA-CSW was designed to adjust the value of the exponent p according to the CS of each gene in S, thus controlling how fast the running-sum statistic increases.

Define the exponent p

Genes with high CS should receive a smaller exponent p to increase the magnitude of the effect whilst the ones with low CS should receive a greater exponent p to reduce the magnitude of their effects. The exponent p is thus defined:

$$p(S,i) = \lambda(1 - CS_i) \tag{3.1}$$

where $p(S,i)$ is the exponent p for the i_{th} hit gene in set S, CS_i is the confidence score of the i_{th} gene. $\lambda \in (0,1]$ (set as 1 as default in the current implementation) is a scaling factor controlling the magnitude of the effect of the CS on the exponent p.

Thus Equation (1.1) can be rewrite as follows:

$$P_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^{\lambda(1-CS_i)}}{N_R}, \text{ where } N_R = \sum_{g_j \in S} |r_j|^{\lambda(1-CS_i)}, P_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{(N - N_H)} \tag{3.2}$$

where i is the position in order list L, r_j is the correlation of gene g_j with class C rescaled to [-1,1] (normalized by maximum), CS_i is the confidence score of gene g_j range from 0 to 1, N is the total number of genes in L and N_H is the number of overlapping gene in S and L.

Since r_j ranges from 0 to 1, the increment step ($P_{hit}(S,i)$) will be bigger for genes with a higher CS value. By adjusting the exponent p, the algorithm rewards the 'good' annotation while penalizes the 'bad' annotations. In this way, if there is a large fraction of 'bad' genes in a particular gene set ranked at one extreme (top or bottom) of the list, the corresponding gene set will have a smaller ES compared to the original implemented GSEA. Note that when all genes in set S have the same confidence score (equally weighted), the above equation will reduce the GSEA-CSW to original GSEA.

The scaling factor λ was used to control the degree of the effect of the confidence score. However, there is no obvious way to determine and optimize the value of λ. No training data is available at the moment to estimate λ. Even if there is training data available, the λ value will be optimized on that particular train data. It is possible that, when multiple independent training data are available, a generic λ could be estimated. Thus, when such training data become available, a better estimation of λ could be learnt than the current implementation of $\lambda = 1$.

Topology method with GSEA-CSW

GSEA-CSW currently supports two topology methods in *topOnto*, *classic* and *elim*. For a given ontology, the *classic* method treats all the nodes in the ontology DAG as a flat list, applies the GSEA-CSW algorithm on each node simultaneously regardless of the DAG hierarchical structure. This is identical to the original *classic* method discussed previously. The *elim* method, on the other hand, investigates the nodes in the DAG level by level from the bottom of the DAG (highest level) to the root. Nodes in the same level are tested simultaneously because there are no edges between them. This way, the algorithm assures that for the currently tested node all children have also been tested. When a node u is found significantly enriched, all of the genes of u are marked as removed in all nodes of $upperInducedGrapgh(u)$, that is in all ancestors of node u. The rationale behind the *elim* method is to try to prevent genes in u repeatedly contributing to the enrichment of u's ancestors. In a count based algorithm like Fisher's exact test, all of the genes in S contribute equally to the enrichment of u so all of them are removed from u's ancestors. However, in a score based approach like GSEA-CSW, it is the leading edge subset (see section 1.2) actually contributing to the enrichment of u while the other genes in u are 'unused', thus removing those 'unused' genes is inappropriate. In order to preserve the effect of the 'unused' genes, options are available to configure the algorithm to remove only the leading edge subset of genes of u, allowing those 'unused' genes to be tested further in the ancestor nodes. Also, due to the 'roll-up' process discussed earlier, not just the genes but their scores are aggregated bottom-up from the ontology DAG. Instead of removing the genes completely, a better alternative to prevent the genes re-contributing to the ancestor nodes is to reduce the gene scores of those genes. These approaches are implemented in *topOnto* and the details of the *elim* GSEA-CSW algorithm is summarized in fig. 3.8.

Algorithm 1 elim GSEA-CSW

```
markedGenes ← ∅ : markedGenesScore ← ∅ : nodeSig ← ∅
elim.type ⊂ ['socre','simple'] : elim.gene.type ⊂ ['core','all']
get the DAG levels list  dagLevels
for i from max(dagLevels) to 1 do                          ▷ iterate from leaves to root
    for u in nodes(dagLevels, i) do
        if elim.type == 'score' then                       ▷ reduce score of previous marked gene in u
            scores[u] ← scores[u] − markedGenesScores[u]
            genes[u] ← genes[u] ∩ (scores[u] ≠ 0)
        else if elim.type == 'simple' then                 ▷ remove previous marked gene from u
            genes[u] ← genes[u] \ markedGenes[u]
        end if
        nodesSig[u] ← GSEACSW(genes[u], scores[u], sigGenes)
        if nodeSig[u] ≤ thresheold then
            for x in upperInducedGraph(u) do               ▷ iterate through all ancestors
                ELIM(u, x, elim.gene.type)
            end for
        end if
    end for
end for
return nodeSig

procedure ELIM(node, target, elim.gene.type)
    if elim.gene.type == 'core' then ▷ mark only the leading-edge subset of genes (core)
        markedGenes[target] ← markedGenes[target] ∪ coregenes[node]
        markedGenesScores[target] ← markedGenesScores[target] + corescores[node]
    else if elim.gene.type == 'all' then                   ▷ mark all the genes in node
        markedGenes[target] ← markedGenes[target] ∪ genes[node]
        markedGenesScores[target] ← markedGenesScores[target] + scores[node]
    end if
end procedure
```

Figure 3.8: The GSEA-CSW algorithm with the elimination topology method.

The GSEA-CSW is particularly suitable for analysis of large lists of genes with gene annotation generated by automatic methods. It is not intended to increase the statistical power of GSEA, but to incorporate important gene annotation score information into the enrichment test process result. Such score can be a reflection of the degree of certainty of the gene disease associations, as well as the source of the gene disease associations. For example, associations from manually curated database are more reliable than others, thus could have a higher weighting in GSEA-CSW algorithm. However, in the currently implementation, the algorithm is blind to the calculation of the confidence score. The actual calculation of the the confidence score, such as those discussed in section 2.3.2, needs to take into account such information.

In terms of validating the method, as pointed out by Huang et al in [16], there is currently no appropriate standard evaluation procedure to evaluate new enrichment methods. The method proposed in [16] aims to evaluate different enrichment methods on a data set generated by randomly shuffling the phenotype labels of an aforementioned experimental data set. The rationale behind this method is that a gene set deemed significantly enriched by more statistical methods is less likely to be false than a gene set deemed significant by fewer statistic methods. An MC (mutual coverage) score is computed for each statistical method against others representing the the degree of mutually identified enriched gene sets between them. This method cannot be directly applied to evaluate the GSEA-CSW due to the lack of standard weighted annotation data. Thus, in the following section, simulated data were used to evaluate the performance of the algorithm.

3.2.7.1 Validation of GESA-CSW with synthetic data

In order to test the algorithm, gene expression data were simulated in a fashion to represent typical gene set structures found in real gene expression data. The objective of the validation is three-fold: 1) to illustrate that GESA-CSW works equal well on structure-less gene set as the original GSEA methods; and 2) to demonstrate GESA-CSW's performance with integrated topology information and 3) to test GESA-CSW's performance on detecting different annotation scores. The three tests were done independently to reduce the complicity of the interpretation of the test result.

Test GESA-CSW against GESA

Gene sets with different levels of differential expression ($\Delta\mu = 0, 0.75$) were generated with varying levels of intra-group correlation ($\rho = 0, 0.6$). Mixed gene sets, i.e. gene

sets that included both differentially expressed and non-differentially expressed genes were also constructed. A flat list of gene sets were used for the original GSEA method, but hierarchically structured in an ontology fashion to be used for GESA-CSW.

The simulated gene expression data set consisted of 1000 genes with $n = 20$ samples, 10 in each of two groups (C1 and C2), for example, disease and control. The data were generated using a 1000-dimensional multivariate normal distribution, with variances set to 1 and means and correlations specified as follows:

- *background:* 960 genes with $\mu = 0 = 0$ and $\rho = 0$.
- *set 1:* (differential expression, correlation): 20 genes with $\Delta\mu = 0.75$ (difference of mean value between two groups, $\mu_{C1} = 0, \mu_{C2} = 0.75$) and pairwise correlation $\rho = 0.6$ among the genes.
- *set 2:* (differential expression, no correlation): same as set 1 but with $\rho = 0$ between the genes in the set.
- *set 3:* (no differential expression, no correlation): 20 genes selected randomly from the background.
- *set 4:* (differential expression, correlation, mixed with background): 10 genes with $\Delta\mu = 0.75$ and $\rho = 0.6$ from set 1, the other 10 randomly selected from the background.
- *set 5:* (differential expression, no correlation, mixed with background): same as set 4 but with $\rho = 0$ between the genes in the set (i.e. 10 genes from set 2 and 10 genes from the background).

Thus, the 1000 simulated genes are constituted by the background set, S1 and S2. The gene set enrichment analysis should be able to detect at least the pure sets S1 and S2, but ideally also the mixed sets S4 and S5 where only half of the genes are differentially expressed. S3 serves as a negative control.

Using this data I first conducted gene set enrichment analysis with the original GSEA algorithm and the GSEA-CSW algorithm with the exponent p set to 1 (ignoring annotation score). The enrichment result is shown in table 3.3. The two methods produced exactly the same result, successfully identified gene set 1,2,4,5 ($p - value < 0.05$) while leaving negative control S3 out. The result indicated that GSEA-CSW is equivalent to the original GSEA method when analysing a flat list of gene sets without any assessment of gene annotation confident within each gene set.

GSEA					GSEA-CSW				
Gene Set	ES	NES	*p*	Group	Gene Set	ES	NES	*p*	GROUP
S1	-0.90918	-1.7937	0	C2	S1	-0.909	-1.794	0.001	C2
S2	-0.85816	-2.0921	0	C2	S2	-0.858	-2.092	0.001	C2
S3	-0.27419	-0.88935	0.8086	-	S3	-0.274	-0.889	0.686	-
S4	-0.68183	-1.7685	0.024	C2	S4	-0.682	-1.769	0.024	C2
S5	-0.74842	-2.0934	0	C2	S5	-0.748	-2.093	0.001	C2

Table 3.3: Enrichment result of simulation data between GSEA and GSEA-CSW. The two methods produce identical result, indicating that GSEA-CSW performed equally good as GSEA with a flat list of gene sets without any annotation confident within each gene set.

Test GESA-CSW with integrated topology information

In order to test GESA-CSW's performance with integrated topology information, four more gene sets, set 6,7,8,9, were generated using the same 1000 simulated genes. The four gene sets are hierarchically structured (fig. 3.9), where S8 is the child set of S9 and the parent set of S6 and S7. The set up are specified as follows:

- *set 6:* 5 genes randomly selected from set 1, 95 genes randomly selected from the background.
- *set 7:* 5 genes randomly selected from set 2, the same 95 genes from background as set 6.
- *set 8:* empty gene set to test the topology effect.
- *set 9:* empty gene set to test the topology effect.

Using the above four gene set, I conducted gene set enrichment analysis using GSEA and GSEA-CSW algorithm with the classic and the elim topology method. The GSEA algorithm does not account for topology information of the gene set, thus the sets are considered as a flat list during the analysis. The exponent p was set to 1 (ignoring annotation score). The enrichment result is shown in table 3.4.

The algorithm behaved as expected for the four gene sets. None of the algorithm pick up set 6 or set 7, which were designed to fail the enrichment test since each of them contains only 5% of differential expressed genes. GSEA did not use any topology information from the sets, thus S8 and S9 were empty during the test. As a result, GSEA does not find any signifiant enriched gene set. In terms of GSEA-CSW, S8 was detected by both topology methods because the annotation roll-up process, that is, the

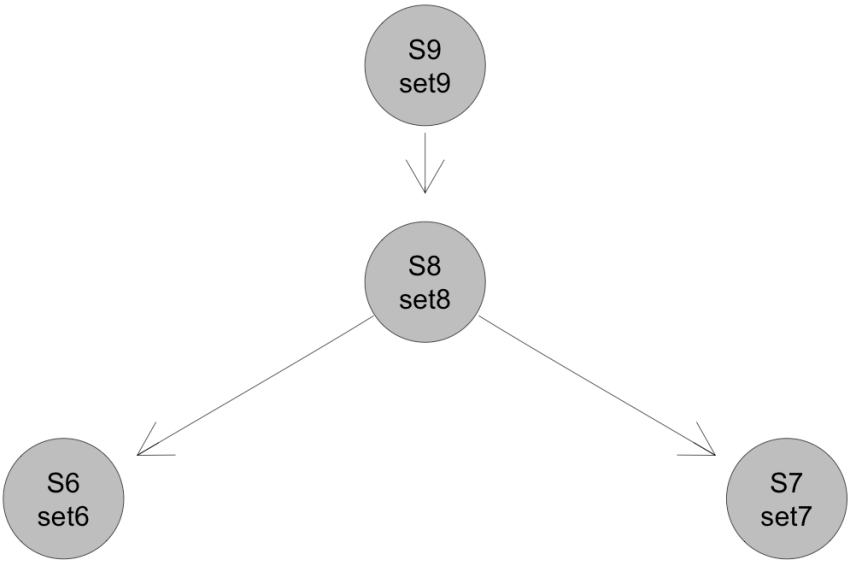


Figure 3.9: The simulated hierarchical structure of the simulated gene set 6,7,8 and 9. S8 is the child set of S9 and the parent set of S6 and S7.

ancestors of S6 and S7 inherited their annotation. The join effect of S6 and S7 were strong enough to be detected when each of them alone was too weak to be detected. S9, however, was assessed as enriched by the classic method, but not the elim method, due to the elimination process which remove the annotation from S9 giving that S8 was found enriched.

GSEA					GSEA-CSW classic					GSEA-CSW elim				
Gene Set	ES	NES	*p*	Group	Gene Set	ES	NES	*p*	Group	Gene Set	ES	NES	*p*	Group
S6	-0.284	-1.248	0.095	C2	S6	-0.284	-1.248	0.095	C2	S6	-0.284	-1.248	0.095	C2
S7	-0.266	-1.171	0.17	C2	S7	-0.266	-1.171	0.17	C2	S7	-0.266	-1.171	0.17	C2
S8	-	-	-	-	S8	-0.329	-1.439	0.014	C2	S8	-0.329	-1.439	0.014	C2
S9	-	-	-	-	S9	-0.329	-1.439	0.014	C2	S9	-	-	-	-

Table 3.4: Enrichment result of simulation data between original GSEA, classic and elim GSEA-CSW. The original GSEA did not find any significant gene set. The GSEA-CSW elim method found S8 enriched while the GSEA-CSW classic methods found S8 and S9 enriched.

Test GESA-CSW with annotation confidence scores

Recall that in GSEA-CSW, an enrichment score (*ES*) is calculated by walking down a ranked gene list L, increasing a running-sum statistic (RES) when encountering a

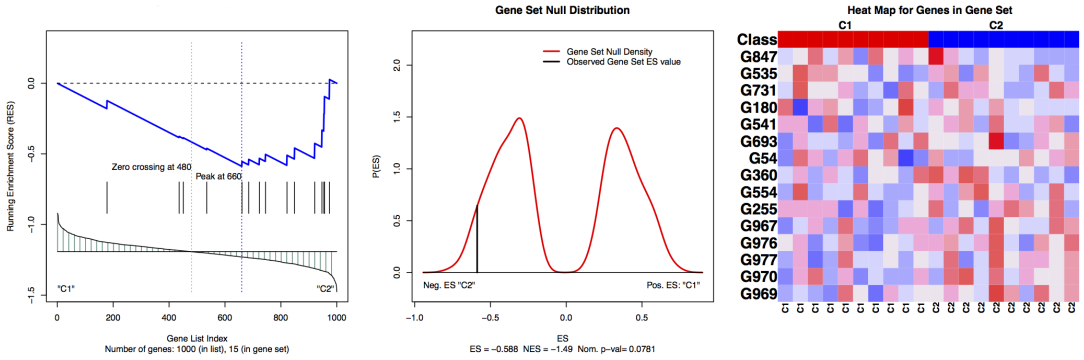
member gene of set S (P_{hit}) and decreasing it when encountering a non-member gene of set S (P_{miss}) (Equation eq. (1.1) on page 18). The magnitude of the increment during the running-sum process is controlled by an exponent p which is calculated base on the annotation confidence score of the gene. Thus, a higher scored gene will result in a larger increment of the RES compared to a lower scored gene, which will also effect on the final enrichment score (ES) and the $p-value$.

In order to illustrate the effect of annotation score in GESA-CSW, a simulated gene set, S10, was generated from the same 1000 simulated genes from above. S10 was designed to have a weak signal, only one third of its genes were differential expressed genes. The objective of this test is to exam the effect of the annotation score in detecting S10, and the corresponding leading edge genes of the gene set. The details were specified as follows:

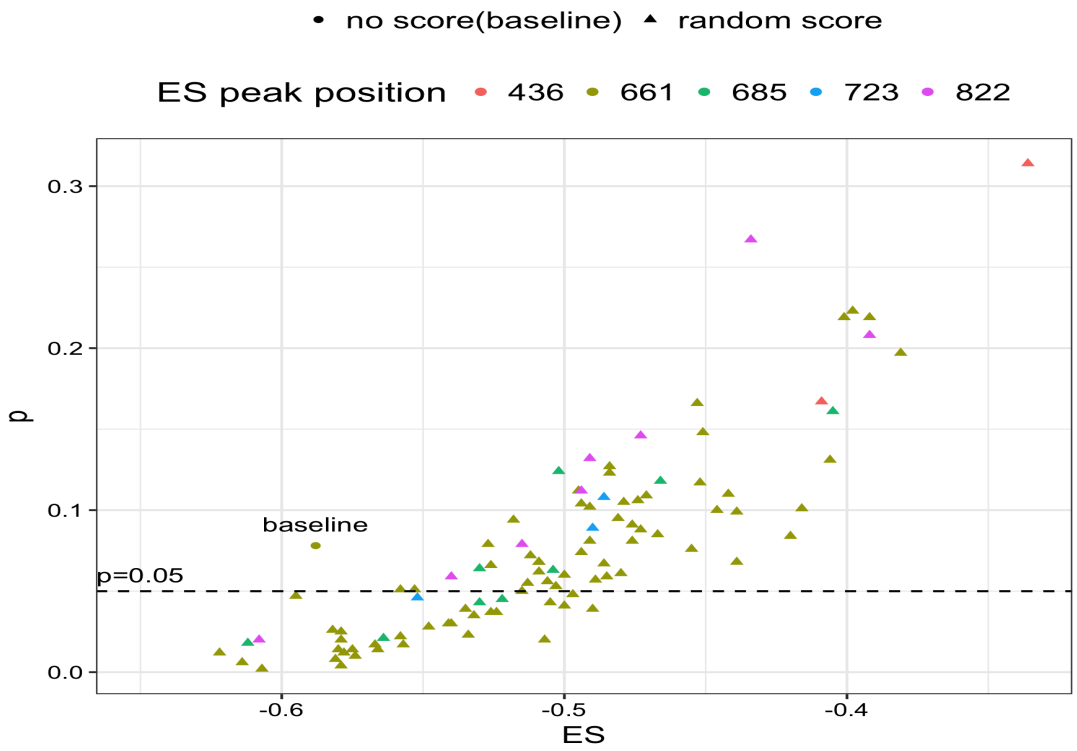
- *set 10:* 5 genes randomly selected from differential expressed genes, 10 genes randomly selected from the background.

S10 was initially assessed without annotation score and was not found significantly enriched($p = 0.078$). The ES score peaked at -0.588 at position 660 in the ranked gene list L as shown in fig. 3.10a. This result was used as a base line to compare with the result using annotation score. Annotation score were randomly generated, between 0 to 1, and assigned to the 15 genes in S10. Such process was repeated 100 times, resulting in 100 different variation of S10, on which GSEA-CSW was performed. The result idicated that the integration of the annotation score affected the ES value its position, which in turn affecting the $p-value$ (fig. 3.10b). This effect of the annotation score can be either positive or negative. i.e, a different annotation score may result in a more extrema ES value and a smaller $p-value$, or it may weaken the enrichment, producing a less extrema ES value and a bigger $p-value$.

Next, annotation scores of genes in S10 were deliberately assigned so that only one of the 15 genes in S10 have a high score (0.9) while the rest have a low score (0.1). The high scored gene was shift among the 15 genes which result in 15 different annotation score variation. GSEA-CSW was then perform and compared to the baseline(S10 without annotation score). The algorithm behave as expected (fig. 3.11). The high scored gene result in a larger increasement of the RES. As the high scored gene shifting from the first gene to the last gene in S10, different maximun ES score appeared in different positons, as expected.



(a)



(b)

Figure 3.10: (a)The enrichment result of gene set 10 with GSEA-CSW without annotation score. S10 was not significantly enriched ($p - value = 0.0781$). (b) Annotation score were randomly assigned the genes in S10 100 times. The enrichment result enrichment result indicated that the ES value, the position of the maximum ES in the ranked gene list, and the $p - value$ vary under different annotation score combination.

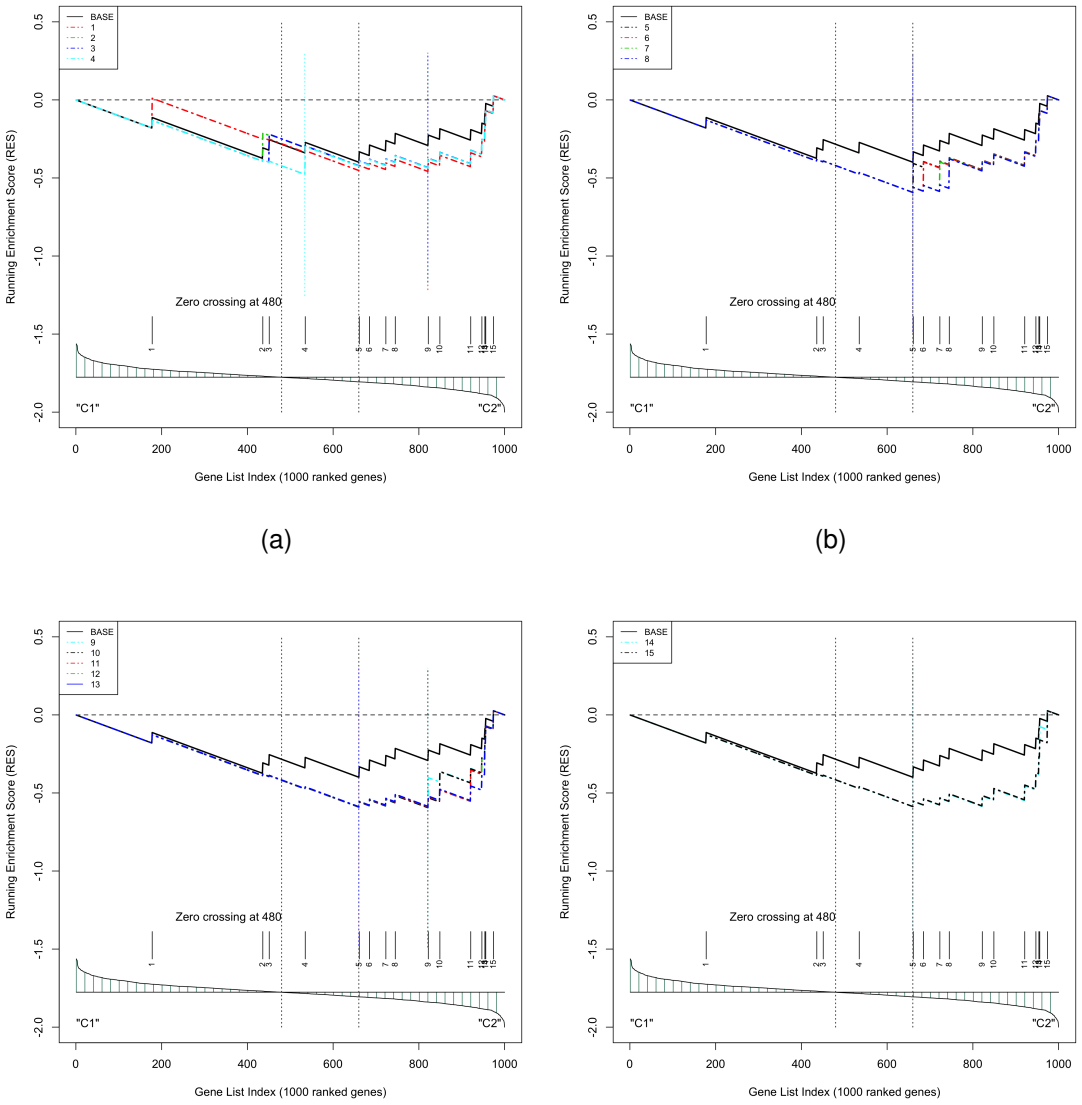


Figure 3.11: Enrichment analysis with GSEA-CSW algorithm on the simulated gene set S10. S10 contains 15 genes(black vertical bar in each graph) located in different position of the 1000 simulated gene rank list L. Annotation scores are assigned to the 15 gene so that each time, one gene has a high score (0.9) while the result are low scored (0.1). Each line of the RES represent one test with the indicated gene being high scored. The BASE (black line) is the result when annotation score was not used. The vertical dash line idicated the position of the maximun ES value. The RES lines vary base on different annotation score set up, indicating the effect of the annotation score in GSEA-CSW algorithm.

3.3 Application of *topOnto*

Arc/Arg3.1 (activity-regulated cytoskeleton-associated protein/activity-regulated gene 3.1), for simplicity henceforth refer to as Arc, is a cytoskeletal protein first characterized in 1995 [274]. It is mainly localized at postsynaptic sites [275] but has also been found in NMDA receptor complexes [276, 277] and PSD95/Dlg4 complexes [278]. Despite Arc having been studied for many years, little is known about its protein complexes or role in disease. Mutation in the human ARC gene has not been directly linked to any mental disorder. However, previous studies show that Arc complexes play an important role in schizophrenia [278, 279] suggesting that Arc complexes may be involved with multiple diseases of cognition. A recent study[5] from Grant et. al. (Seth.Grant@ed.ac.uk) identified, in mice, 107 high-confidence Arc interactors from the Arc complexes. Human genetic studies identified mutations and variants in Arc interacting proteins that are enriched in schizophrenia, intellectual disability, epilepsy and normal variation in intelligence. To reaffirm and extend the evidence of the Arc complexes role in disease, enrichment analysis was performed with *topOnto* using the Human Disease Ontology (HDO), Human Phenotype Ontology, Gene Ontology (GOBP, GOCC, GOMF) and Reactome Pathway Ontology (RPO). The 107 protein encoding genes in mice are projected to 106 human homology genes, which we use as input for enrichment analysis, with one missing gene B630019K06Rik (NCBI Entrez id:102941) could not be mapped from mouse to human. Fisher's exact test was used with four topology methods *classic,elim,weight01* and *parentchild*. The top 10 enriched terms for each topology method with HDO are shown in table 3.5. The corresponding results for the other ontologies are list as appendix in the table A1-table A5 on page 226.

HDO The enriched HDO terms for genes from the ARC complex make biological sense (table 3.5). *Schizophrenia* is the most significantly enriched disease found by *classic,elim* and *weight01* methods while its parent terms '*psychotic disorder, cognitive disorder*' and '*disease of mental health*' were also identified by *classic* and *parentchild* methods. Neurodegenerative diseases like '*Parkinson's disease*' and '*Alzheimer's disease*', brain diseases including '*epilepsy, frontotemporal dementia*' and '*cerebral degeneration*', and cognitive disorders like '*vascular dementia*' were also found enriched. Other enriched diseases not shown here including '*autistic disorder*' (elim

[5] https://www.genes2cognition.org/publications/tap-arc/

TERM.ID	Term	Level	classic	elim	weight01	parentchild
DOID:5419	schizophrenia	6	7.8e-21	7.8e-21	9.4e-20	0.77349
DOID:2468	psychotic disorder	5	9.8e-21	1.00000	1.00000	7.8e-06
DOID:936	brain disease	6	1.0e-17	0.10353	1.00000	1.1e-07
DOID:1561	cognitive disorder	4	1.6e-16	0.66238	1.00000	0.00223
DOID:150	disease of mental health	3	1.8e-16	0.09385	1.00000	1.8e-16
DOID:1826	epilepsy syndrome	7	4.1e-15	0.00020	3.6e-05	0.00091
DOID:331	central nervous system disease	5	3.4e-12	0.46574	1.00000	0.00293
DOID:863	nervous system disease	4	9.8e-11	0.11586	1.00000	2.3e-07
DOID:1443	cerebral degeneration	7	5.4e-10	5.4e-10	3.6e-11	0.00242
DOID:9255	frontotemporal dementia	8	1.0e-09	1.0e-09	1.0e-09	0.24237

(a) classic

TERM.ID	Term	Level	classic	elim	weight01	parentchild
DOID:5419	schizophrenia	6	7.8e-21	7.8e-21	9.4e-20	0.77349
DOID:1443	cerebral degeneration	7	5.4e-10	5.4e-10	3.6e-11	0.00242
DOID:9255	frontotemporal dementia	8	1.0e-09	1.0e-09	1.0e-09	0.24237
DOID:3328	temporal lobe epilepsy	9	7.8e-07	7.8e-07	7.8e-07	0.29681
DOID:84	osteochondritis dissecans	8	5.1e-06	5.1e-06	5.1e-06	0.01030
DOID:8725	vascular dementia	6	1.1e-05	1.1e-05	1.1e-05	0.00552
DOID:14330	Parkinson's disease	8	1.2e-05	1.2e-05	4.2e-05	0.42649
DOID:0060125	heavy chain disease	7	2.7e-05	2.7e-05	0.03239	0.00079
DOID:10652	Alzheimer's disease	8	7.5e-05	7.5e-05	7.5e-05	0.91964
DOID:1826	epilepsy syndrome	7	4.1e-15	0.00020	3.6e-05	0.00091

(b) elim

TERM.ID	Term	Level	classic	elim	weight01	parentchild
DOID:5419	schizophrenia	6	7.8e-21	7.8e-21	9.4e-20	0.77349
DOID:1443	cerebral degeneration	7	5.4e-10	5.4e-10	3.6e-11	0.00242
DOID:9255	frontotemporal dementia	8	1.0e-09	1.0e-09	1.0e-09	0.24237
DOID:3328	temporal lobe epilepsy	9	7.8e-07	7.8e-07	7.8e-07	0.29681
DOID:84	osteochondritis dissecans	8	5.1e-06	5.1e-06	5.1e-06	0.01030
DOID:8725	vascular dementia	6	1.1e-05	1.1e-05	1.1e-05	0.00552
DOID:1826	epilepsy syndrome	7	4.1e-15	0.00020	3.6e-05	0.00091
DOID:14330	Parkinson's disease	8	1.2e-05	1.2e-05	4.2e-05	0.42649
DOID:10652	Alzheimer's disease	8	7.5e-05	7.5e-05	7.5e-05	0.91964
DOID:0050709	Ohtahara syndrome	10	0.00035	0.00035	0.00035	0.59558

(c) weight01

TERM.ID	Term	Level	classic	elim	weight01	parentchild
DOID:150	disease of mental health	3	1.8e-16	0.09385	1.00000	1.8e-16
DOID:936	brain disease	6	1.0e-17	0.10353	1.00000	1.1e-07
DOID:863	nervous system disease	4	9.8e-11	0.11586	1.00000	2.3e-07
DOID:2468	psychotic disorder	5	9.8e-21	1.00000	1.00000	7.8e-06
DOID:7	disease of anatomical entity	3	8.5e-05	0.27073	1.00000	8.5e-05
DOID:0080008	ischemic bone disease	7	0.00014	0.30680	1.00000	0.00034
DOID:0060125	heavy chain disease	7	2.7e-05	2.7e-05	0.03239	0.00079
DOID:1826	epilepsy syndrome	7	4.1e-15	0.00020	3.6e-05	0.00091
DOID:1561	cognitive disorder	4	1.6e-16	0.66238	1.00000	0.00223
DOID:1443	cerebral degeneration	7	5.4e-10	5.4e-10	3.6e-11	0.00242

(d) parentchild

Table 3.5: Disease enrichment analysis of the ARC complex with different topology methods. The resulting enriched disease terms vary based on the topology method used.

p=4.8$e-3$), '*intellectual disability*' (elim p=2.5$e-2$) and one of its particular subtypes '*autosomal dominant non-syndromic intellectual disability*' (elim p=1.2$e-3$). There is a possible error in the enrichment result - '*osteochondritis dissecans*' which is type of bone disease is enriched that could be error in the annotation. This term is annotated with 21 genes (42 times in total, 35 from GeneRIF, 1 from OMIM, 6 from Var). By checking the annotation, only 5 genes (10 times) are correctly annotated to this term. The other 16 genes (32 times) are wrongly annotated to this term because of the inappropriate usage of the ambiguous synonym, 'OCD', in the HDO. The abbreviation 'OCD' can refer to osteochondritis dissecans and obsessive-compulsive disorder at the same time, but only appears as synonym in the former term definition. As a result, OntoSuite-Miner wrongly linked genes that are found relevant with obsessive-compulsive disorder to osteochondritis dissecans. The use of abbreviation is a frequent source of error which is discuss in section 2.3.4.1 on page 76. These types of error are induced by the ambiguous abbreviation synonym defined in the ontology, suggesting that ontologies should be carefully used distinguishing terminology to avoid mapping problems, especially with abbreviations.

HPO Similar to the result from HDO, *Schizophrenia* was also the most significantly enriched term in HPO (table A1). '*Seizures*', '*Alzheimer disease*', '*Dementia*' and terms from '*Encephalopathy*' (brain disease, damage, or malfunction) were also found to be enriched. '*Autism*' (elim p=5.3$e-3$), '*Intellectual disability*' (elim p=2.4$e-3$) and one of its child terms *Intellectual disability, severe* (elim p=2$e-2$) are reported to be significant but not shown here.

ReactomePathway The most enriched pathway for the ARC complexes is *Unblocking of NMDA receptor, glutamate binding and activation* (table A2). When a neuron is not sending a signal, it is "at rest". The NMDA receptor is blocked by extracellular Mg^{2+} ions and is not activated in this state by ligands. This pathway involves removing the block from the NMDA receptor which activates it. Pathways involved in the neurotransmitter release cycle, such as the '*glutamate neurotransmitter release cycle*' and '*GABA synthesis, release, reuptake and degradation*', receptor binding and recycling including '*Insulin receptor recycling*', '*Ion channel transport*' were also found to be enriched. Downstream signaling pathways including '*Trafficking of AMPA receptors*', '*post NMDA receptor activation events*' and its downstream '*Ras activation event*' were reported to be enriched. Besides all the signaling pathways, another in-

teresting pathway, '*Prefoldin mediated transfer of substrate to CCT/TriC*', is enriched. Prefoldins are a family of proteins working as a transfer proteins in conjunction with the chaperonins, CCT/TriC, to form a chaperone complex and correctly fold other proteins. One of prefoldin's main uses is the formation of molecules of actin for use in the cytoskeleton. This enrichment reflects the functional role of ARC complex in the cytoskeleton by regulating the folding and formation of actins and tubulins.

GO The Gene Ontology Biological Process describes genes and their functions (table A3). '*synaptic transmission*' is the most significantly enriched term while other functional terms relevant to ion transport, cell-cell signaling and localization such as '*sodium ion export from cell*' and '*ionotropic glutamate receptor signaling pathway*' were also enriched. Interestingly, the term '*learning*' (Any process in an organism in which a relatively long-lasting adaptive behavioral change occurs as the result of experience) is enriched which agrees with the HDO and HPO results, suggesting that the ARC complexes play an important role in memory and experience-related behave patterns. Enrichment with the Gene Ontology Cellular Component reveals the localization of ARC complexes. Postsynaptic structures like '*postsynaptic membrane, postsynaptic density*' and ὲextitdendritic spine' were enriched indicating that the ARC complex is mainly localized at postsynaptic sites. The enrichment of '*endocytic vesicle membrane, synaptic vesicle*' and '*extracellular exosome*' support observations in [280, 281] that the ARC complex interacts with the endocytic machinery to regulate AMPA receptor trafficking. In addition, '*NMDA selective glutamate receptor complex*' (not shown, elim p=$3.3e-05$) was found to be enriched, backing up previous findings from [276, 277]. Further evidence from the Gene Ontology Molecular Function shows that '*extracellular-glutamate-gated ion channel activity, sodium:potassium-exchanging ATPase activity, NMDA glutamate receptor activity*' and '*AMPA glutamate receptor activity*' (not shown, elim p=$2.3e-4$) were enriched emphasizing again the important role of ARC complexes in the endocytic machinery.

By summarizing enrichment results across different ontologies, a better and more complete picture of the ARC complex is achieved. The complex is mainly localized in postsynaptic sites which contain extensive molecular machinery that link the postsynaptic membranes and presynaptic membranes together and carry out the signaling process. Different types of receptor such as NMDARs and AMPARs are located at the postsynaptic membranes which receive neurotransmitter, glutamate for example, and transfer them into the cell body to trigger downstream signaling cascades. The

ARC complex also exists in receptor complexes and plays an important role in activation and regulation of these receptors. Evidence shows that the genes in the ARC complexes are significantly associated with mental health diseases and nervous system diseases including schizophrenia, Alzheimer's disease, Parkinson's disease, intellectual disability, autism, and epilepsy. Among these diseases, schizophrenia shows the strongest link to the genes found in the ARC complexes.

Generally speaking, the *classic* and *parentchild* approaches tend to identify terms that are close to the root and provide a top-down view of the gene set in relation to the ontologies. For example, the HDO terms '*cognitive disorder*' and '*nervous system disease*' are very general terms identified by the *classic* and *parentchild* methods, but not by *elim* and *weight01*. On the other hand, *elim* and *weight01* tend to find more specific terms such as '*Alzheimer's disease*', which was not reported to be significant by the *parentchild* method. The average depth of the top 10 enriched terms for different topology methods across different ontologies are shown in fig. 3.12a and different performance of the topology methods are shown in fig. 3.12b. It is recommended to look at the result from the *classic* and *parentchild* to get an overview of the gene set in relation to the ontologies first, then move to the result of the *elim* and *weight01* for more specific information.

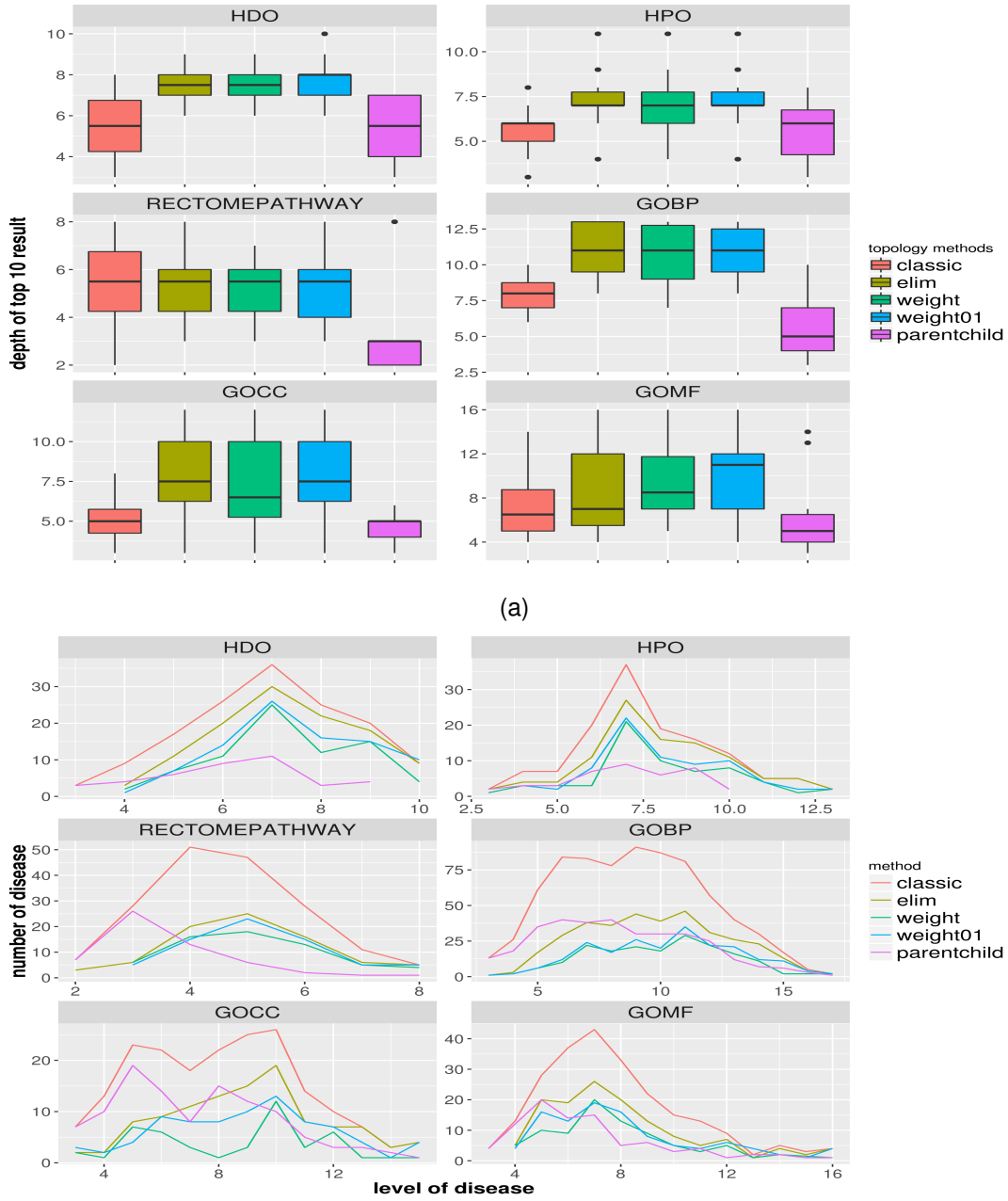


Figure 3.12: Depth of (a)the top 10 enriched terms and (b) all enriched terms for the ARC complex with different topology methods. The *classic* and *parentchild* methods tend to identify terms that are close to the root and provide a top-down view. While the *elim* and *weight01* methods usually identify specific terms that are close to the leaves of the ontology DAG, thus provide a bottom-up view of the ontology DAG. The *classic* method generated the largest number of enriched diseases among all the methods.

3.4 Conclusions and future work

In this chapter I reviewed the usefulness and common limitations of current enrichment tools, and proposed the second part of the *OntoSuite* framework, the *OntoSuite-Analytics*, which consists of a set of R packages including *topOnto* and the corresponding data packages for the ontologies. I discussed statistical limitations, back-end annotation database, multiple hypothesis correction and the key design decisions to ameliorate them. The usage of the package was illustrated by analysing the activity-regulated cytoskeleton-associated protein (ARC) complex.

The *topOnto* package was built on top of the existing bioconductor *topGO* package, which is specifically designed for GO enrichment analysis. It implements a range of popular statistical algorithms and topology methods for gene set enrichment analysis and facilitates easy cross comparison of enrichment results between different statistic/-topology combinations. *topOnto* provides all the features of the *topGO* package with extra features including 1) An mechanism to run enrichment analysis across any standard ontology and present the result for easy comparison, 2) addition methods to work on the ontology DAG and 3) the addition of the GSEA and the GSEA-CSW methods. One of the key features *topOnto* provides is unified enrichment analysis across ontologies. Linux shell scripts are available to convert ontologies stored in standard OBO format, into the corresponding ontology package that can be used by *topOnto*. The package currently supports the HDO (Human Disease Ontology), HPO (Human Phenotype Ontology), PCO (Panther Protein Class Ontology), CO (Chromosome Ontology), RPO (Reactome Pathway Ontology) and the three GO ontologies GO-BP (Biological Process), GO-MF (Molecular Function) and GO-CC (Cellular Component). Scripts are available to convert ontologies from a standard OBO file into the corresponding ontology package that to be used by *topOnto*. Currently in September 2016, 106 ontologies are represented natively in OBO in the NCBO bioportal ontology repository while 351 ontologies are represented natively in OWL. OBO has became a sub language of OWL, and can be created from OWL. Therefore, potentially there are more than 450 ontologies that can be used in enrichment analysis by *topOnto*. Since NCBO has been the central repository of biomedical ontologies, it would be interesting to build a pipeline that pragmatically accesses ontologies from the NCBO and converts them into ontology packages that can be used by *topOnto*.

A new algorithm, named GSEA-CSW was implemented in *topOnto* as a modification of the original GSEA algorithm, which is particularly suitable for enrichment anal-

ysis when the back-end annotation data are scored. GSEA-CSW differs from GSEA on the way it controls the step length when calculate a Kolmogorov-Smirnov like statistic. GSEA-CSW takes into account the gene confidence score, thus amplifying the contribution of high scored gene associations in the enrichment analysis.

The GSEA-CSW algorithm was described and implemented and tested with simulated data but not with real biological data.. This is because evaluating the performance of enrichment methods is difficult due the absence of any gold standard, especially for weighted association which has not been done before. [16], Huang et.al developed a system to evaluate different enrichment methods by computing an MC (mutual coverage) score, representing the degree of mutually identified enriched gene set between them. Thus enrichment methods with a higher MC score are suggested to perform better than others. This system cannot be directly applied to evaluate the GSEA-CSW due to the lack of standard weighted annotation data. Further evaluation of the algorithm is needed in the future with the availability of standard evaluation methods and testing data.

The GSEA method is computationally intensive due to the permutation process, and so is GSEA-CSW. This problem is amplified when the algorithm is implemented into *topOnto* when the *elim* method is used taking into account the ontology structure. This results in a long processing time when using GSEA-CSW with the *elim* topology method. The *elim* algorithm considers one level of ontology terms at a time and the result will affect the next level of terms, thus within one level, the algorithm is parallelalizable. However, the cost of dividing/distributing the task, and merging the result needs to be calculated and tested further to prove that it is worth implementing.

The *topOnto* package, together with the gene disease association data set *HDGDB* discussed in chapter 2 on page 33, have been used by Mclean et.al. in [282], a co publication with me as the second author, to validate the performance of a novel scalable modularity based clustering algorithm, the Spectral Modularity algorithm, which can be use to detect community structure in biological networks. *topOnto* package was used in the study to assess the disease relevance of the detected communities in three previously studied biological networks including 1)the MASC complex, representing a protein complex surrounding the mammalian NMDA receptor [283] consists of 101 proteins and 246 interactions, 2) the PostSynaptic Density (PSD) [284] consists of 1312 proteins and 8031 protein interactions, and 3) the Human interactome network BioPlex [285], which contains 7668 proteins and 23744 protein interactions, found using high-thoughput affinity purification mass spectrometry in human embry-

onic kidney (HEK) 293T cells. The disease enrichment result was used to evaluate the performance of different community detection algorithms with some unexpected implication of the connection between the BioPlex network and Alzheimers disease.

topOnto is implemented as an R package using the R/Bioconductor system and is freely available on github (https://github.com/statbio/topOnto). The package has been submitted to Bioconductor and is being reviewed at the time of writing this thesis. It would be much more useful to implement it as a user-friendly web interface like DAVID [53] and DisGeNET [192].

Chapter 4

A comprehensive disease profile of human genes

The search for feature enrichment in gene sets is a widely used characterisation method. Instead of focusing on individual genes, such analyses try to summarize the information for a set of genes grouped together based on shared features. For example, genes that are involved in the same pathway or genes that located in a certain region of the chromosome. *HDGDB* contains high-quality gene annotations for human diseases. It should be of great utility for the exploration of relationships between and within gene sets in human disease research using *HDGDB*.

Ontology annotation provides a natural way to categorize human genes. Genes annotated to the same ontology term are implied to be involved in the same biological concepts. In the following section, I use four ontologies to categorize genes including 1) genes involved in the same pathways based on the Reactome database [172] (Reactome pathway ontology, RPO), 2) gene products belonging to the same protein classes based on Panther Protein Class [173, 174] (Protein class ontology,PCO), 3) genes that are located in the same chromosome region (limited to only one sub-band level) of the human genome (Chromosome ontology,CO) and 4) genes that are associated with the same human diseases (Human disease ontology, HDO). Note that strictly speaking, RPO, PCO and CO are not formally defined ontologies. They are however terminologies that contain all essential ontology features including unique and consistent ids of terms, names, definitions and most importantly, the relations between terms. These terminologies are structured hierarchically so that the terms near the root of the structure are more generic, exactly as for an ontology. Even though these terminologies do not always provide all the optional information that is commonly found in an ontology,

for example, cross links to other ontology terms, the existing data provided by the terminologies and has covered all the essential features in an ontology thus, is sufficient to be used in this way. As a result, the above three terminologies were parsed into ontologies and subsequently used in the enrichment analysis.

4.1 Profiling gene sets with disease based annotations

HDGDB contains high quality gene annotations for human disease, making it possible to construct a comprehensive disease profile for gene sets. Three ontologies, including RPO, PCO and CO and their gene annotations were used to construct the gene sets of interests, for example, genes that are located in the same chromosome region or their encoded proteins are involved in the same pathway. A gene set, in terms of the enrichment test, is conceptually similar to an ontology term, where genes are grouped by their characteristic. It is interesting to use all the terms in RPO (1921), PCO (245) and CO (114) to construct the gene sets and all of the terms in the HDO (6819) for disease enrichment analysis. This results in a large amount of gene sets and generates very detailed results for each of them. However, it is sometimes more interesting to summarize disease information, for example for all signaling molecules (PC00207), rather a profiling a specific type of signaling molecule, for example, Chemokine (PC00074) which are a family of small signaling proteins that exert their biological effects by interacting with G protein-linked transmembrane receptors called chemokine receptors [286].

In order to do this, a set of generalized terms like 'PC00207 signaling molecules' needs to be selected from the ontologies to give a broad overview of the ontology content without the detail of the specific fine grained terms. However, ontologies were defined independently by different consortium, and have different levels of detail in each domain. In an ontology DAG, the leaf node represents the most specific concept in a branch. The number of leaf nodes in a certain level in the ontology varies between ontologies (fig. 4.1). Some of the leaf nodes are as deep as level 14 (14 nodes away from the ontology root) while others are much closer to the root. The unbalance of ontology structures makes it difficult to select general terms from any given ontology. Note that CO is an exception because it represents chromosome regions detailed to one sub band level, thus has a maximum of 3 levels and the level 3 terms are those representing the individual sub bands.

Manually selection of terms can be accurate but requires domain knowledge and is subjective since it is unclear how to sensibly compare between terms derived from

different ontologies (for example: 'PC00207 signaling molecules' in PCO versus 'R-HSA-162582 Signal Transduction' in RPO) and even within the same ontology ('PC00194 protein kinase receptor' versus 'PC00193 protein kinase' and 'PC00197 receptor'). Another approach to select a set of general terms is by using the ontologies' structures themselves, choosing an appropriate level of abstraction based on the ontologies' hierarchy. This requires minimum domain knowledge and reflects the native parent-children relationship between ontology terms. However, ontologies were created to represent different domain knowledge, thus have different structures. A uniform level cut will result in the terms representing different level of abstraction among ontologies. To avoid subjective selection of the ontology terms, I used the latter approach and manually inspected the ontologies in order to select a proper level. As a result, I choosed the level 3 terms of RPO ,PCO and CO to build disease profiles, even though this approach is still suffered from the ontology unbalanced structure. For the same reason, 136 level 3 HDO terms were used in order to generate a reasonable degree of granularity to represent all human disease. This resulted in a total of 277 ontology terms and 73 PCO terms, 128 RPO terms and 76 CO terms being profiled. PCO[1], RPO[2] and CO[3] annotations were taken from the corresponding websites on 20^{th} July 2016. Gene annotations were rolled-up before the profiling process using the 'true path rule' of the ontology definition where all attributes (including gene annotation) of the children must hold for all parents. Thus, for example, the level 2 RPO term 'Apoptosis' has 35 children including 'Intrinsic Pathway for Apoptosis', 'Regulation of Apoptosis' and 'a poptotic execution phase'. Genes annotated to these 35 terms, were aggregated into the term 'Apoptosis'. Disease enrichment analysis was then perform on each of the 290 constructed gene sets (ontology terms) against the 136 HDO terms using *topOnto* and *HDGDB* with the *elim* topology methods. During enrichment analysis, the gene background (the reference gene set used in Fisher type significant test) was defined as the overlapping genes between a particular ontology and HDO. For example, we profiling RPO terms, genes that do not exist in RPO annotation but present in HDO annotation are removed. This is similar to the enrichment analysis performed on microarray data where the gene background is usually defined as all the genes measured by the microarray.

The ontology terms tested were grouped by their parent terms. For example, in

[1] http://www.pantherdb.org/downloads/
[2] http://www.reactome.org/pages/download-data/
[3] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/

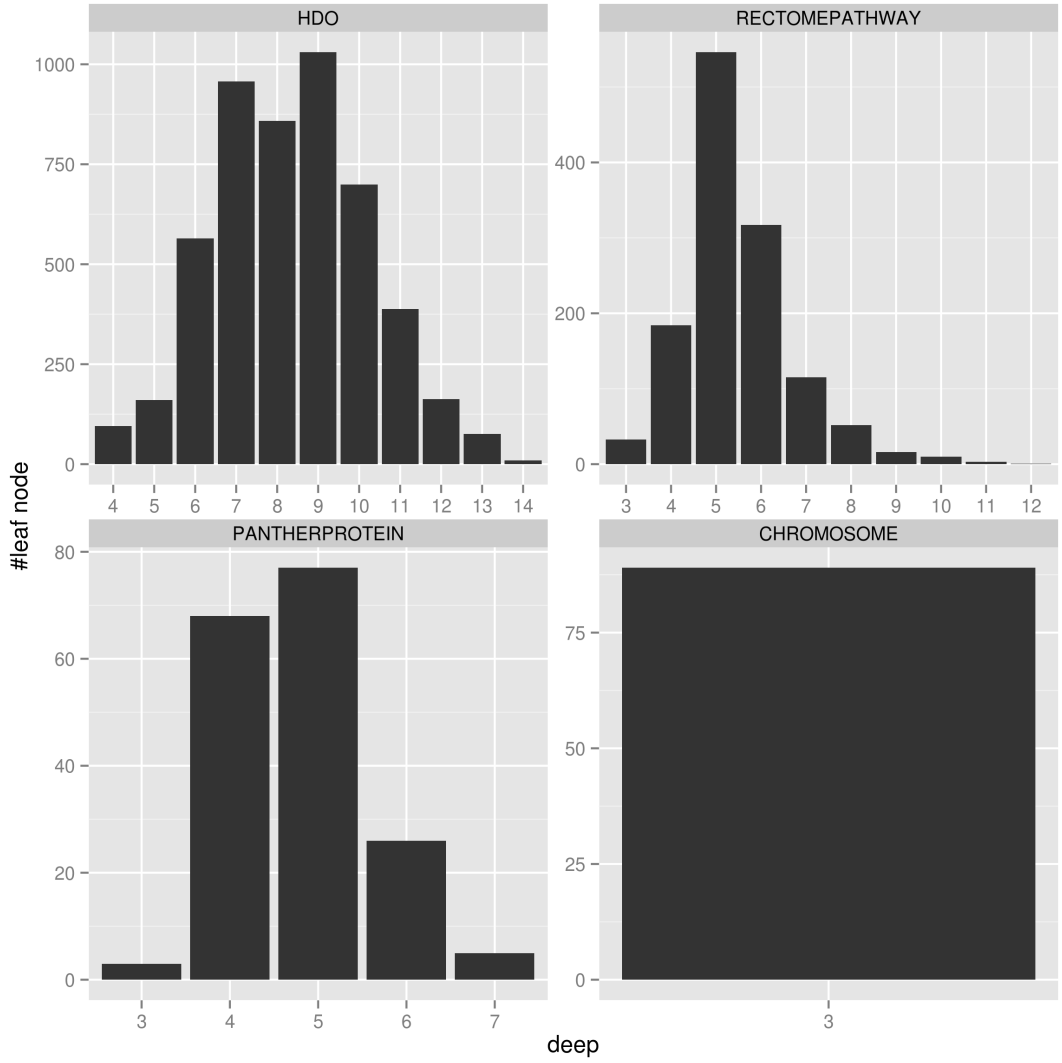


Figure 4.1: The number of leaf nodes in a certain level in the ontology varies between ontologies. Some of the leaf nodes are as deep as level 14 (14 nodes away from the ontology root) while other is little closer as 2 nodes away from the root.

the CO, terms representing sub bands including '1p1', '1p2', '1p3', '1q2', '1q3' and '1q4' were grouped with the parent CO term 'chromosome_1'; while in RPO, a series of signaling pathways such as 'Signalling by Activin' and 'Signalling by NOTCH' were grouped into the parent RPO term 'Signal Transduction'. In order to explore consistent patterns from the disease profile within each group, it would be interesting to find out those diseases that were found enriched across multiple terms within the same group. Let's define N_G as the number of ontology term tested in group G, and $N_{D,G}$ as the number of times a disease D was found to be enriched within group G,

thus:

$$P_{D,G} = \frac{N_{D,G}}{N_G} * 100\% \tag{4.1}$$

A high $P_{D,G}$ value indicates that disease D is consistently enriched across group G, such a pattern may be of interested for further exploration. For example, the pathway group 'R-HSA-73894 DNA Repair' contains 6 pathways, out of which 5 were enriched for 'DOID:162 cancer'. $P_{DOID:162,R-HSA-73894} = 83\%$. In addition to the enrichment analysis, a $P_{D,G}$ was calculated for each disease-group pair across ontologies. Those groups that contain only one term are ignored since the $P_{D,G}$ is always 100% in this case. For example, pathway group 'R-HSA-1474165 Reproduction' contains only one sub pathway 'R-HSA-1187000 Fertilization'.

The following of this section details the setup of the disease enrichment process and presents the results from each of the three ontologies including CO, RPO and PCO, will be discussed. Some interesting observations from the results are discussed further with supporting evidence from the literature.

4.1.1 Chromosome region

Each human chromosome has a short (denoted 'p) and long arm (denoted 'q'), separated by a centromere. Each chromosome arm is further divided into regions, or cytogenetic bands, that can be seen using a microscope and special stains. The cytogenetic bands are labeled p1, p2, q1, q2, etc., counting from the centromere out toward the ends of the chromosome which are called telomeres. At higher resolutions, sub-bands can be seen within the bands. The sub-bands are also numbered from the centromere out towards the telomere. For example, the cytogenetic map location of the CFTR gene is 7q31.2, which indicates it is on chromosome 7, q arm, band 3, sub-band 1, and sub-sub-band 2.

The non-random distribution of polymorphic loci associated with breast cancer [287] and schizophrenia [288] was reported and suggested that human chromosomes are heterogeneous in structure and function. Identifying the correlation between human diseases and chromosome location can provide useful etiological insights and help prioritize likely causal relationships. Disease enrichment analysis is an ideal approach to explore such relationships and the results provide a human disease 'hot-spot' profile for each of the 76 chromosome sub-band locations tested and can be used to answer questions such as 'what is the most enriched disease in chromosome 22q1?'. The Chromosome Ontology (CO) was created to represent chromosome location based

on chromosome structure and is used here in enrichment analysis. Gene annotations were downloaded from the EntrezGene database and aggregated to sub-band level, which contains 37638 human genes annotated with 76 CO terms. The summarized results are shown in fig. A1. Previous studies support many of the results. For example, 'DOID:1561 cognitive disorder' was found enriched in 13 chromosome sub bands, among which 22q1 (p=$9.7e-7$) has the most significant result. This is due to 22q11.2 deletion syndrome, which is a disorder caused by the deletion of a small piece of chromosome 22q11.2, and which has been linked to developmental delays, including delayed growth and speech development, and learning disabilities. People with 22q11.2 deletion syndrome also have an increased risk of developing mental illnesses such as schizophrenia, depression, anxiety, and bipolar disorder later in life [289,290]. Another example is 'DOID:2914 immune system disease', which is defined as *a disease of anatomical entity that is located in the immune system* in the Human disease ontology, that was found to be significantly enriched (top 5 elim p$<=0.05$) in 8 chromosome sub-bands including 1q2 (p=$1.47e-2$), 1q3 (p=$4.9e-9$), 6p1 (p=$4.3e-2$), 6p2 (p=$7e-13$), 6q2 (p=$2.8e-2$), 9p2 (p=$9.8e-3$), 12q1 (p=$2.6e-3$) and 18q2 (p=$1.5e-2$). In particular, the most dense area was found on 1q3, where more than 20% of genes were linked to some type of immune system disease, followed by 6p2 and 18q2 with the corresponding gene percentages being 14% and 12.6% respectively. The top enriched immune system disease (not shown here) in 1q3 is 'DOID:12554 hemolytic-uremic syndrome' (p=$6.5e-15$), which is a condition caused by the abnormal destruction of red blood cells clogging the filtering system in the kidneys [291]. Other diseases including 'DOID:10608 celiac disease' (p=$5e-4$) and 'DOID:626 complement deficiency' (p=$5.7e-4$) were also found enriched [292, 293]. Previous studies have confirmed the strong relationship between chromosome 6 and immune system disease. The HLA gene family [294] including more than 220 different genes are located close together on a 3 Mbp stretch within chromosome 6p2, and provides instructions for making a group of related proteins known as the human leukocyte antigen (HLA) complex. The HLA complex helps the immune system distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria; its function is essential to the immune system. Diseases including 'DOID:8857 lupus erythematosus' (p=$2.7e-28$) where human immune system becomes hyperactive and attacks healthy tissues, and 'DOID:12361 Graves' disease' (p=$3.9e-18$) which is the leading cause of hyperthyroidism, a condition in which the thyroid gland produces excessive hormones, are among the top enriched diseases in 6p2 [295].

On average, 40% of the genes on chromosome sub bands are annotated with at least one disease according to our *HDGDB* data (fig. 4.2). The most highly disease annotated sub band is 16q2 where 57.3% genes were annotated with at least one disease, while Yq1 is the least annotated sub band, only 3.7% of its genes were found associated with any disease. Most disease genes are annotated to a small number of diseases. Others (mostly cancer related genes) are linked to a large number of diseases such as TNF (427) on 6p2, VEGFA (414) on 6p1 and TP53 (401 disease) on 17p1.

A weak positive linear relationship (Pearson correlation = 0.29) was observed between the number of genes located on a chromosome sub band region and the number of enriched diseases found in that region (fig. 4.3). Some of the regions contain a large number of genes but show little relationship in terms of diseases (1p3, 19q1, 11q1, ect.) while other smaller regions show stronger relationship with diseases (3p2, 9p2, 8p2, ect.). In particular, 1p3 is the largest chromosome sub band containing 1335 genes, but only two types of disease were found to be significantly enriched, 'DOID:0060239 Van der Woude syndrome' (p=$2.3e-5$) and 'DOID:535 sleep disorder' (p=$2.7e-02$). 19q1 is the second largest chromosome sub band with 1182 genes, but only one disease 'DOID:8545 malignant hyperthermia' (p=$2.4e-2$) was found enriched. On the other hand, 9p2 is a relatively small sub band with 287 genes while 11 types of diseases were found enriched including 'DOID:863 nervous system disease' (p=$1.1e-4$),'DOID:934 viral infectious disease' (p=$6.9e-4$) and 'DOID:1561 cognitive disorder' (p=$6e-3$). 6p2, containing 1003 genes, was identified as the most dense sub band for disease where 19 types of diseases were enriched.

To explore the disease pattern of each chromosome, the $P_{D,G}$ value was calculated and those with $P_{D,G} > 50\%$ are shown in table 4.1. It is observed that three types of diseases, 'DOID:863 nervous system disease' with chromosome 10, 'DOID:0060037 developmental disorder of mental health' with chromosome X and 'DOID:15 reproductive system disease' with chromosome Y, were found enriched across the whole chromosome group which suggests a close relationship between these diseases and the corresponding chromosome region. Some of these relationships had been observed in previous literature [296–301].

The disease profile of chromosomes could indicate the importance of chromosome regions in a disease context. It also provides a set of dense areas in chromosomes for particular type of diseases, which may be used as a guidance to discover novel gene functions or gene disease associations. It is observed that the chromosome area where known chromosomal rearrangements (truncation, deletion. etc.) occur are more

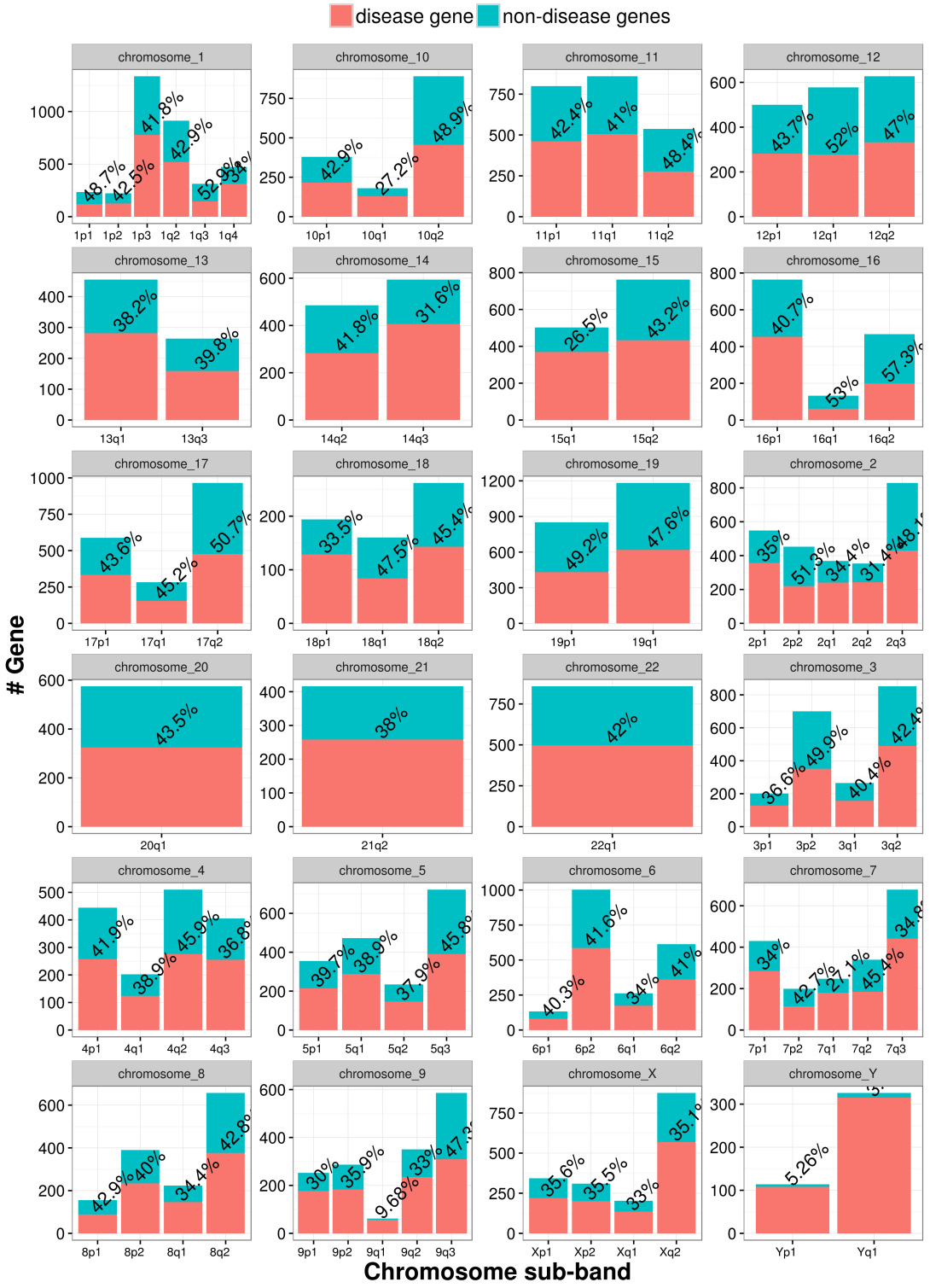


Figure 4.2: The distribution of disease and non disease genes on each chromosome sub band. On average, 40% of the genes in chromosome sub bands are annotated with at least one disease according to *HDGDB* data. The most disease annotated sub band is 16q2 with 57.3% of genes annotated with at least one disease, while Yq1 is the least annotated sub band with only 3.7% of genes associated with disease.

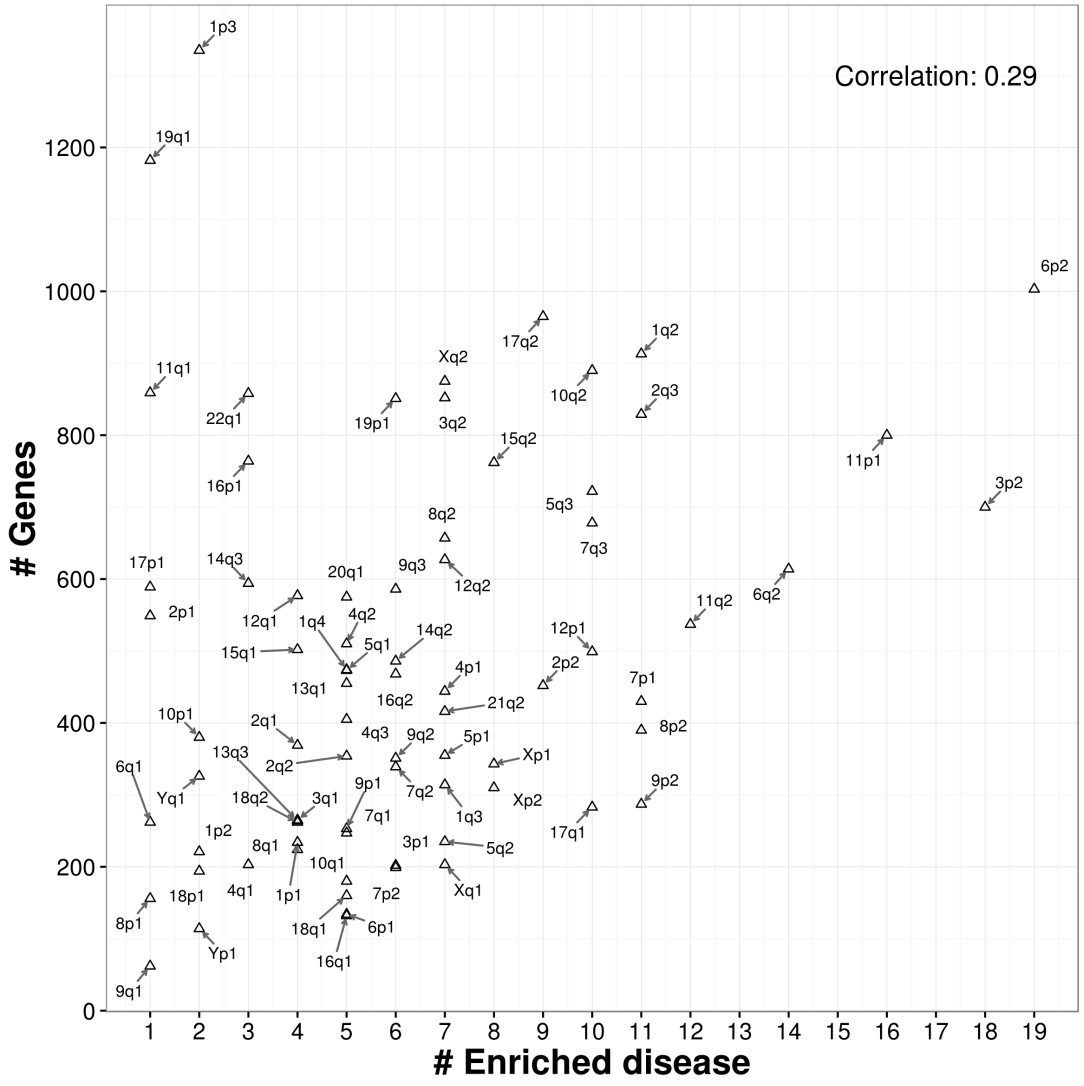


Figure 4.3: Chromosome size and the number of enriched diseases. A weak positive linear relationship (Pearson correlation = 0.29) was observed between the number of genes located on a chromosome sub band region and the number of enriched diseases found in that region

Group	Term.ID	Term.DEF	$N_{D,G}$	N_G	$P_{D,G}$
chromosome_2	DOID:7	disease of anatomical entity	3	5	60.0%
chromosome_5	DOID:1579	respiratory system disease	3	4	75.0%
chromosome_6	DOID:2914	immune system disease	3	4	75.0%
chromosome_6	DOID:1287	cardiovascular system disease	3	4	75.0%
chromosome_9	DOID:0060158	acquired metabolic disease	3	5	60.0%
chromosome_10	DOID:1579	respiratory system disease	2	3	66.7%
chromosome_10	DOID:863	nervous system disease	3	3	100.0%
chromosome_11	DOID:1398	parasitic infectious disease	2	3	66.7%
chromosome_11	DOID:9007	sudden infant death syndrome	2	3	66.7%
chromosome_12	DOID:0060158	acquired metabolic disease	2	3	66.7%
chromosome_12	DOID:17	musculoskeletal system disease	2	3	66.7%
chromosome_12	DOID:0014667	disease of metabolism	2	3	66.7%
chromosome_16	DOID:104	bacterial infectious disease	2	3	66.7%
chromosome_16	DOID:0060340	ciliopathy	2	3	66.7%
chromosome_16	DOID:0050545	visceral heterotaxy	2	3	66.7%
chromosome_17	DOID:16	integumentary system disease	2	3	66.7%
chromosome_18	DOID:1287	cardiovascular system disease	2	3	66.7%
chromosome_X	DOID:0050177	monogenic disease	3	4	75.0%
chromosome_X	DOID:0060037	developmental disorder of mental health	4	4	100.0%
chromosome_Y	DOID:15	reproductive system disease	2	2	100.0%

Table 4.1: Chromosome disease pair with $P_{D,G} > 50\%$.

likely to become disease hot spots, for example 22q1 [302]. However, these result may be biased by the uneven research focus on genes, i.e, some genes (regions) are more studied than others, and are therefore better annotated, which increases the likelihood finding enriched diseases. Further data and tests are needed to distinguish between the above two possibilities for each particular chromosome region.

4.1.2 Reactome pathway

A biological pathway is a series of actions among molecules (proteins, small molecules, genes, etc) in a cell that leads to a certain product or a change in the cell. For example, a pathway can trigger the assembly of new molecules, turn genes on and off, or spur a cell to move. There are many types of biological pathways. Among the most well-known are pathways involved in metabolism, in the regulation of genes and in the transmission of signals. In the Reactome pathway database, large pathways were divided into sub-pathways, which may be further divided into smaller sub-pathways

until reaching the individual steps in a pathway, known as reactions. There are 1921 pathways/reactions in Reactome (20 July 2016) and these are structured hierarchically (fig. 4.4). From top to bottom of the pathway hierarchy, pathways become smaller, more specific, and finally divide into reactions.

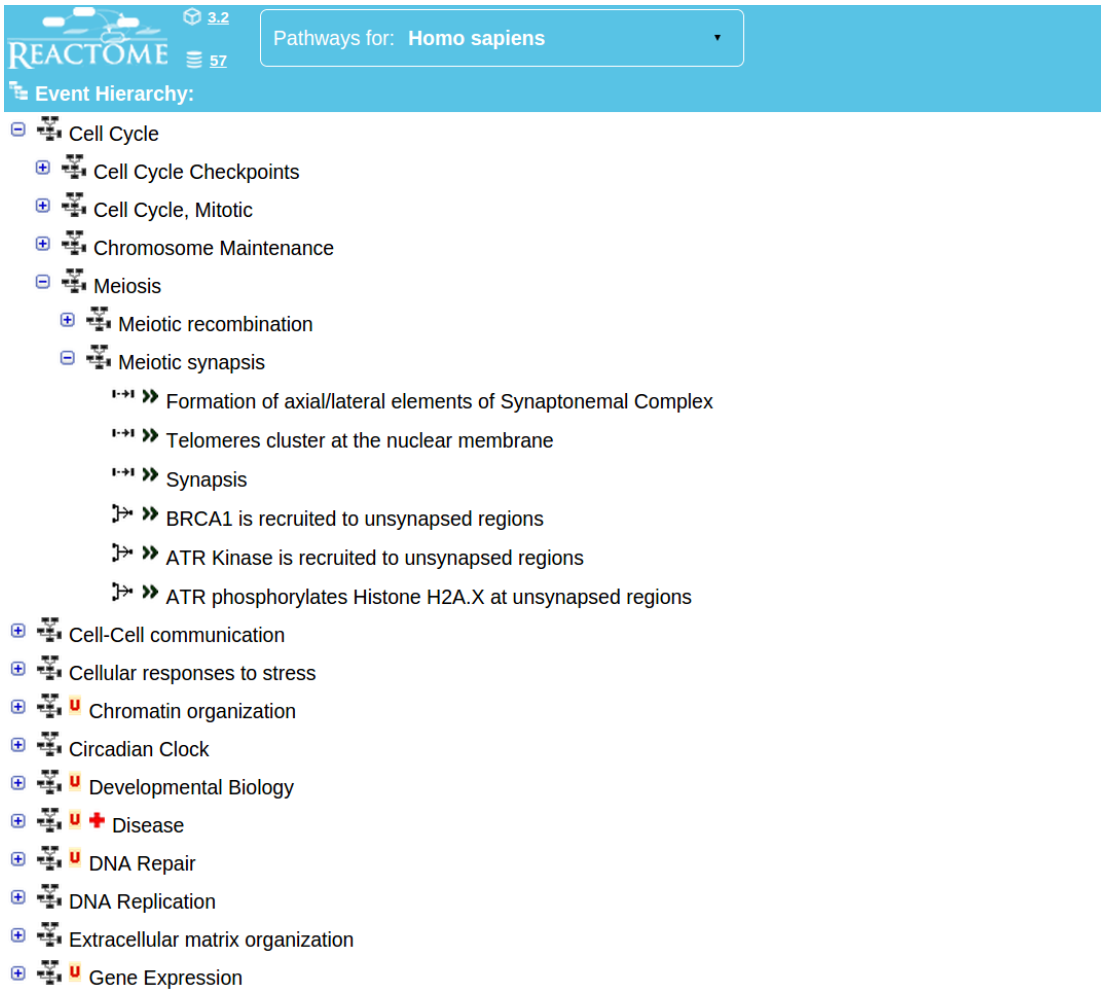


Figure 4.4: Reactome pathway hierarchical structure. From top to bottom of the pathway hierarchy, large pathways were divided into sub-pathways, which may be further divided into smaller sub-pathways until reaching the individual steps in a pathway, known as reactions.

In order to profile Reactome pathway with human diseases, Reactome pathway and the pathway annotation data were downloaded from the Reactome website on 17/02/2016. An ontology, refer to as RPO (Reactome pathway ontology), was created to represent Reactome pathway based on the pathway hierarchy and subsequently used in the enrichment analysis. Pathway annotations, containing 8446 unique human genes, were aggregated to 145 level 3 RPO terms. belonging to 24 level 2 Reactome

pathway, which were subsequently used in disease enrichment analysis with *topOnto* and *HDGDB*. *elim* topology methods were used to estimate the enrichment p-value. As a result, 128 Reactome pathways were found with at least one enriched disease. The summarized results are shown in fig. A2. The $P_{D,G}$ value for each pathway-disease pair was calculated and those with $P_{D,G} > 50\%$ are shown in table 4.2.

The results make biological sense and most of them can be confirmed by reviewing the literature. 'R-HSA-109582 Hemostasis' (fig. A2c) is a physiological response that culminates in the arrest of bleeding from an injured vessel. The HDO term 'DOID:1287 cardiovascular system disease' was found enriched with $P_{DOID:1287,Hemostasis} = 85.7\%$. This disease includes a set of blood, heart, blood vessel and lymphatic system disease that are closely related to crucial processes like clot formation, platelet activation and interactions with the vascular wall. In Reactome pathway group 'R-HSA-397014 Muscle contraction', it is observed that closely related diseases such as 'DOID:1287 cardiovascular system disease' ($P_{D,G} = 100\%$) and 'DOID:17 musculoskeletal system disease' ($P_{D,G} = 67\%$) were found enriched.

The 'R-HSA-112316 Neuronal System' (fig. A2a) represents neurons in the brain and their communication mechanisms. The communication occurs across synapses, the functional connection between neurons. Neurotransmitters which are chemical agents released by presynaptic neurons, are transmitted to postsynaptic neurons and activate specific receptor molecules. Neurons' physiological changes in the brain (neuronal loss, changes in the synaptic physiology) results in variable degrees of cognitive decline which result in a wide ranger of nervous system disease and cognitive disorders [303].

The Reactome pathway group 'R-HSA-168256 Immune System' (fig. A2b) represents the ability of the human body to avoid/defend infection other organisms. Viral/bacterial/parasitic infectious diseases, as well as 'DOID:2914 immune system disease' were significantly enriched in this group. Interestingly, 'DOID:162 cancer' was enriched with 3 out of 4 ($P_{D,G} = 75\%$) Reactome pathways including 'R-HSA-1280218 Adaptive Immune System' (p=$6.4e-10$), ' R-HSA-168249 Innate Immune System' (p=$1.8e-16$) and 'R-HSA-1280215 Cytokine Signaling in Immune system' (p=$1.9e-16$). Until recently, investigations into the nature of cancer focused strictly on the cancer cell and on cancer as a genetic disease. This is perfectly illustrated by the six consensus characteristics (hallmarks), proposed by Hanahan and Weinberg [304], to define cancerous cells including the capacity to sustain proliferative signaling, to resist cell death, to induce angiogenesis, to enable replicative immortality, to activate in-

vasion and metastasis, and to avoid growth suppressors. These hallmarks are reflected by the enrichment results of cancer in Reactome pathway groups including 'R-HSA-162582 Signal Transduction' ($P_{D,G} = 83\%$), 'R-HSA-2262752 Cellular responses to stress' ($P_{D,G} = 80\%$), ' R-HSA-73894 DNA Repair' ($P_{D,G} = 83\%$), 'R-HSA-1640170 Cell Cycle' and 'R-HSA-5357801 Programmed Cell Death' ($P_{D,G} = 100\%$). However, a picture of cancer is emerging with a new understanding of the relationship between immunology and cancer. Immunology was previously considered not to be a critical discipline for understanding cancer is now providing important new clues to cancer biology. Four additional hallmarks were proposed, out of which two of them highlight the newly recognized dual interaction between cancer and the immune system: the ability to avoid immune destruction which results in acute inflammation and cancer elimination, and the potential for chronic inflammation that promotes tumor growth rather than elimination [305, 306]. Studies of cancer and immune system interactions have revealed that every known innate and adaptive immune effector mechanism participates in tumor recognition and control [307], and immunotherapy, which works on strengthening the cancer patient's immune system by improving its ability to recognize the tumor or providing a missing immune effector function, is one of the recent cancer treatment approaches that holds promise of a life-long cure [308].

In 'R-HSA-1852241 Organelle biogenesis and maintenance' (fig. A2d), a serious of cilia related diseases, refereed to as Ciliopathies which are human diseases arising from disruption of cilia structure and/or function, were enriched in the Reactome pathway 'R-HSA-5617833 Assembly of the primary cilium'. The primary (nonmotile) cilia are closely related to sensory and cell signaling functions. Availability of low-cost next generation sequencing has facilitated the explosion of new knowledge in the genetic etiology of ciliopathies and reviewed that many genes are shared in common between otherwise clinically distinct ciliopathies [309]. 'DOID:1148 polydactyly' (p=$5.1e-14$) and 'DOID:2490 congenital nervous system abnormality' (p=$1.9e-5$), which have been recently linked to cilia defect [310,311], as well as the master term for ciliopathy 'DOID:0060340 ciliopathy' (p=$1e-30$)' were found significantly enriched.

'R-HSA-4839726 Chromatin organization' (fig. A2e) refers to the composition and conformation of complexes (chromatin) between DNA, protein, and RNA. The complexes decrease the accessibility of DNA but also help to protect it from damage. The most enriched type of disease in this group is 'DOID:162 cancer' (p=$4.4e-06$). In fact, one of the first steps in cancer is the acquisition of genome instability which can be the result of changes in chromatin structure. Polak et.al. [312] compared the ge-

nomic distribution of mutations of 173 cancer genomes from eight different cancer types, to 424 epigenetic features and observed that epigenomics features indicative of active chromatin and transcription were associated with low mutation density, whereas repressive chromatin features were associated with regions of high mutation density. It was concluded that chromatin accessibility and modification, together with replication timing, explain up to 86% of the variance in mutation rates along cancer genomes tested in the paper.

'R-HSA-5205647 Mitophagy' contains only one sub Reactome pathway 'R-HSA-5205685 Pink_Parkin Mediated Mitophagy' (fig. A2g). This Reactome pathway is the process of selective removal of damaged mitochondria by autophagosomes which is induced by PINK1 activities [313]. PINK1 causes the parkin protein to bind to depolarized mitochondria to induce autophagy of those mitochondria. Mutation in the PINK1 gene leads the aggregation of mitochondria which has been associated with Parkinson's disease [314]. Its regulatory roles for mitochondrial function also suggests that it is possibly involved in the pathogenesis of other nervous system disease such as 'DOID:10652 Alzhermer's disease' [315] and 'DOID:12217 lewy body disease' [316], which reflects the enriched HDO term 'DOID:863 nervous system disease' ($p=1.2e-3$) in this group.

'R-HSA-1430728 Metabolism' (fig. A2h) describes processes in human cells that generate energy and mediate the synthesis of diverse essential molecules, as well as the inactivation and elimination of toxic ones generated endogenously or present in the extracellular environment. 'DOID:0080074 neural tube defect', which is mainly caused by folic acid deficiency [317], was found enriched in 'R-HSA-196854 Metabolism of vitamins and cofactors' ($1.5e-10$) and 'R-HSA-211859 Biological oxidations' ($1.5e-5$), while 'DOID:655 inherited metabolic disorder', which in the vast majority of cases is the result of mutations in TCA cycle enzymes, was also enriched in the group ($P_{D,G} = 53.8\%$).

'R-HSA-162582 Signal Transduction' (fig. A2i) represents the process in which extracellular signals elicit changes in cell state and activity. Transmembrane receptors sense changes in the cellular environment by binding ligands, such as hormones and growth factors, or reacting to other types of stimuli, such as light. Such stimulation of transmembrane receptors leads to their conformational change which propagates the signal to the intracellular environment by activating downstream signaling cascades. Depending on the cellular context, this may impact cellular proliferation, differentiation, and survival. At the organism level, signal transduction regulates overall growth

and behavior. 'DOID:162 cancer' was enriched in this group of Reactome pathways with $P_{D,G} = 83.3\%$ indicating its close relationship with cell signaling. Some of the Reactome pathways including 'R-HSA-5683057 MAPK family signaling cascades' (65 cancer terms), 'R-HSA-195721 Signaling by Wnt' (67) and 'R-HSA-157118 Signaling by NOTCH' (68), together with growth factor related pathway including 'R-HSA-190236 FGFR' (69), 'R-HSA-1236394 EGFR' (70) and 'R-HSA-194138 VEGF' (72) were found linked to a large number of types of cancers, pinpointing potential commonalities in tumour signalling mechanisms.. In fact, signal transduction has long been used to explain cancer mechanisms and as a therapeutic target [318–321]. The dy-regularization of signaling pathways often leads to changes in the tumor microenvironment, angiogenesis, inflammation and in gene expression and cellular metabolism, which result in the inactivation of tumor suppressor genes that normally ensure that cells do not proliferate inappropriately or survive outside their normal cycle. For example, the Receptor tyrosine kinase (RTK) pathways including EGFR [322], FGFR [323], insulin receptor [324], NGF [325], PDGF [326] and VEGF [327], frequently activate downstream signaling through RAF/MAP kinases [109, 328], AKT [329] and PLC-gamma [330], which ultimately results in tumor suppressors being inhibited such as p16 [331]. The Hippo pathway plays a critical role in regulating contact inhibition of proliferation [332]. The disruption of this pathway suppresses the transcriptional coactivator YAP, which is emerging as a key tumor suppressor pathway in many cancers [333–335]. The hyperactivated NOTCH pathway can stimulate the cell cycle and also inhibits apoptosis in T cells which contributes to cancer such as acute lymphocytic leukemia [336]. The involvement of Notch signaling in many cancers has led to investigation of notch inhibitors (especially gamma-secretase inhibitors) as potential cancer treatments [337].

The disease profile of Reactome pathways presents a board overview of pathways in the context of human disease. On average, each Reactome pathway was annotated with 151 genes and was found enriched with 12 diseases. A weak positive linear relationship (Pearson correlation = 0.41) was observed between the number of genes in a Reactome pathway and the number of enriched diseases found within that pathway (fig. 4.5). Various signalling transduction pathways were crowded on the bottom right part of the figure, indicating their close relationship with a variety of diseases. These pathways were often responsible for controlling cell proliferation and regulating cell growth, thus tightly related to a variety of cancers. On the other hand, some other pathways are less versatile in terms of disease and only a few diseases were enriched.

For example, only one disease 'DOID:15 reproductive system disease' (p=3.1$e-3$) was found enriched in the 'R-HSA-1187000 Fertilization' pathway.

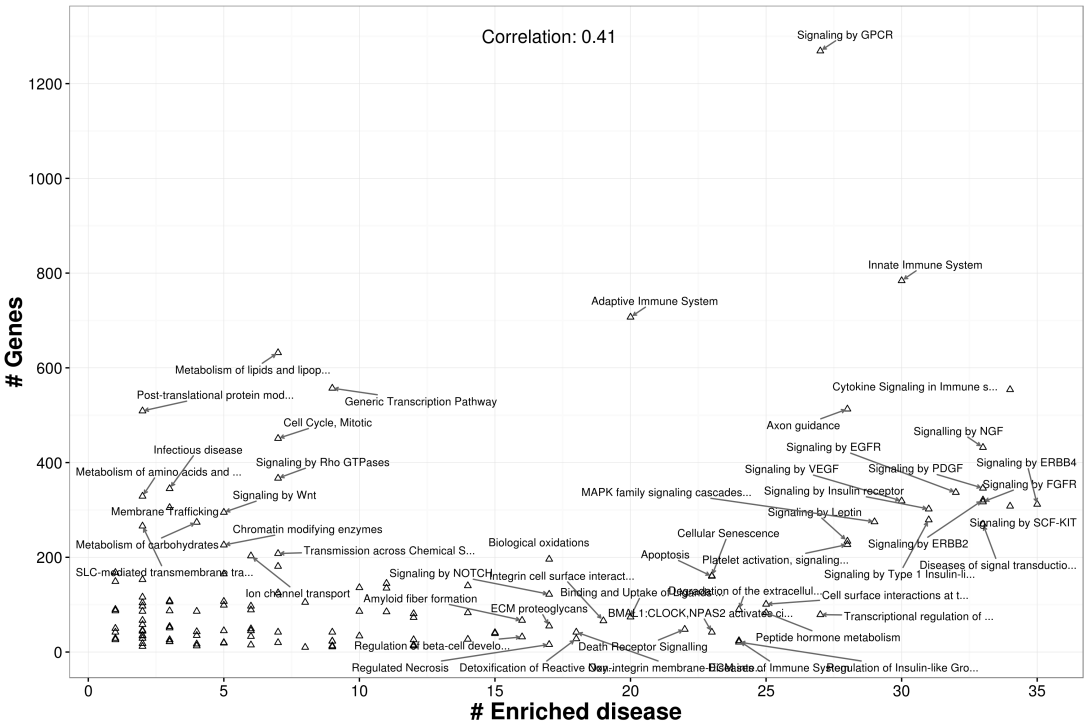


Figure 4.5: The correlation between the size of the Reactome Pathway and the number of significantly enriched disease. A weak positive linear relationship (Pearson correlation = 0.41) was observed between the number of genes in a Reactome pathway and the number of enriched diseases found within that pathway.

A surprisingly high number of genes (average 81% across all pathways) were annotated with Reactome pathways are disease genes that were linked to at least one HDO term. Almost all genes in the pathway group 'R-HSA-1474244 Extracellular matrix organization' were disease genes (fig. 4.6). This is much higher than the corresponding figure among Chromosome genes because Reactome pathways are well defined molecular networks which when disturbed, usually result in a cascade of changes of other downstream pathways often leading to disease.

Group	Term.ID	Term.DEF	N_{D,G}	N_{G}	P_{D,G}
Hemostasis	DOID:77	gastrointestinal system disease	4	7	57.1%
Hemostasis	DOID:18	urinary system disease	4	7	57.1%
Hemostasis	DOID:934	viral infectious disease	4	7	57.1%
Hemostasis	DOID:15	reproductive system disease	4	7	57.1%
Hemostasis	DOID:0060043	sexual disorder	4	7	57.1%

Continued on next page

Table 4.2 – continued from previous page

Group	Term.ID	Term.DEF	N_{D,G}	N_{G}	P_{D,G}
Hemostasis	DOID:1398	parasitic infectious disease	4	7	57.1%
Hemostasis	DOID:162	cancer	5	7	71.4%
Hemostasis	DOID:17	musculoskeletal system disease	5	7	71.4%
Hemostasis	DOID:1579	respiratory system disease	5	7	71.4%
Hemostasis	DOID:2914	immune system disease	6	7	85.7%
Hemostasis	DOID:1287	cardiovascular system disease	6	7	85.7%
Neuronal System	DOID:1561	cognitive disorder	2	2	100.0%
Neuronal System	DOID:0060037	developmental disorder of mental health	2	2	100.0%
Neuronal System	DOID:863	nervous system disease	2	2	100.0%
Neuronal System	DOID:9007	sudden infant death syndrome	2	2	100.0%
Developmental Biology	DOID:162	cancer	4	6	66.7%
Developmental Biology	DOID:0060072	benign neoplasm	5	6	83.3%
Metabolism	DOID:655	inherited metabolic disorder	7	13	53.8%
Extracellular matrix organization	DOID:2914	immune system disease	4	7	57.1%
Extracellular matrix organization	DOID:934	viral infectious disease	4	7	57.1%
Extracellular matrix organization	DOID:0050177	monogenic disease	4	7	57.1%
Extracellular matrix organization	DOID:9250	acrocallosal syndrome	4	7	57.1%
Extracellular matrix organization	DOID:162	cancer	6	7	85.7%
Extracellular matrix organization	DOID:77	gastrointestinal system disease	6	7	85.7%
Extracellular matrix organization	DOID:18	urinary system disease	6	7	85.7%
Extracellular matrix organization	DOID:1287	cardiovascular system disease	6	7	85.7%
Extracellular matrix organization	DOID:15	reproductive system disease	6	7	85.7%
Extracellular matrix organization	DOID:1579	respiratory system disease	6	7	85.7%
Extracellular matrix organization	DOID:16	integumentary system disease	7	7	100.0%
Extracellular matrix organization	DOID:17	musculoskeletal system disease	7	7	100.0%
Extracellular matrix organization	DOID:0060072	benign neoplasm	7	7	100.0%
Signal Transduction	DOID:0060071	pre-malignant neoplasm	13	24	54.2%
Signal Transduction	DOID:0060233	cardiofaciocutaneous syndrome	13	24	54.2%
Signal Transduction	DOID:303	substance-related disorder	14	24	58.3%
Signal Transduction	DOID:0080014	chromosomal disease	14	24	58.3%
Signal Transduction	DOID:2914	immune system disease	15	24	62.5%
Signal Transduction	DOID:934	viral infectious disease	15	24	62.5%
Signal Transduction	DOID:863	nervous system disease	15	24	62.5%
Signal Transduction	DOID:28	endocrine system disease	15	24	62.5%
Signal Transduction	DOID:0050567	orofacial cleft	15	24	62.5%
Signal Transduction	DOID:1579	respiratory system disease	15	24	62.5%
Signal Transduction	DOID:77	gastrointestinal system disease	16	24	66.7%
Signal Transduction	DOID:17	musculoskeletal system disease	16	24	66.7%
Signal Transduction	DOID:15	reproductive system disease	16	24	66.7%
Signal Transduction	DOID:16	integumentary system disease	17	24	70.8%
Signal Transduction	DOID:0050177	monogenic disease	17	24	70.8%
Signal Transduction	DOID:1287	cardiovascular system disease	18	24	75.0%
Signal Transduction	DOID:0060072	benign neoplasm	19	24	79.2%
Signal Transduction	DOID:162	cancer	20	24	83.3%
Cell Cycle	DOID:162	cancer	3	4	75.0%
Cell Cycle	DOID:2490	congenital nervous system abnormality	3	4	75.0%
Cell Cycle	DOID:0060071	pre-malignant neoplasm	4	4	100.0%
Immune System	DOID:162	cancer	3	4	75.0%
Immune System	DOID:934	viral infectious disease	3	4	75.0%
Immune System	DOID:0060043	sexual disorder	3	4	75.0%

Continued on next page

Table 4.2 – continued from previous page

Group	Term.ID	Term.DEF	N_{D,G}	N_{G}	P_{D,G}
Immune System	DOID:77	gastrointestinal system disease	4	4	100.0%
Immune System	DOID:2914	immune system disease	4	4	100.0%
Immune System	DOID:17	musculoskeletal system disease	4	4	100.0%
Immune System	DOID:104	bacterial infectious disease	4	4	100.0%
Immune System	DOID:1398	parasitic infectious disease	4	4	100.0%
Immune System	DOID:1579	respiratory system disease	4	4	100.0%
Cellular responses to stress	DOID:77	gastrointestinal system disease	3	5	60.0%
Cellular responses to stress	DOID:16	integumentary system disease	3	5	60.0%
Cellular responses to stress	DOID:18	urinary system disease	3	5	60.0%
Cellular responses to stress	DOID:0060072	benign neoplasm	3	5	60.0%
Cellular responses to stress	DOID:1287	cardiovascular system disease	3	5	60.0%
Cellular responses to stress	DOID:162	cancer	4	5	80.0%
Muscle contraction	DOID:17	musculoskeletal system disease	2	3	66.7%
Muscle contraction	DOID:1561	cognitive disorder	2	3	66.7%
Muscle contraction	DOID:150	disease of mental health	2	3	66.7%
Muscle contraction	DOID:1287	cardiovascular system disease	3	3	100.0%
Circadian Clock	DOID:162	cancer	2	2	100.0%
Circadian Clock	DOID:934	viral infectious disease	2	2	100.0%
Circadian Clock	DOID:1561	cognitive disorder	2	2	100.0%
Circadian Clock	DOID:0060158	acquired metabolic disease	2	2	100.0%
Circadian Clock	DOID:11612	polycystic ovary syndrome	2	2	100.0%
Circadian Clock	DOID:0060043	sexual disorder	2	2	100.0%
Programmed Cell Death	DOID:162	cancer	2	2	100.0%
Programmed Cell Death	DOID:77	gastrointestinal system disease	2	2	100.0%
Programmed Cell Death	DOID:2914	immune system disease	2	2	100.0%
Programmed Cell Death	DOID:16	integumentary system disease	2	2	100.0%
Programmed Cell Death	DOID:17	musculoskeletal system disease	2	2	100.0%
Programmed Cell Death	DOID:18	urinary system disease	2	2	100.0%
Programmed Cell Death	DOID:934	viral infectious disease	2	2	100.0%
DNA Repair	DOID:0050177	monogenic disease	4	6	66.7%
DNA Repair	DOID:162	cancer	5	6	83.3%

Table 4.2: Pathway disease pairs with $P_{D,G} > 50\%$.

4.1.3 Panther protein class

PANTHER (Protein Analysis Through Evolutionary Relationships) protein classification system (version 7.0) contains 243 protein classes ranging from general terms including 'PC00095 enzyme modulator' and 'PC00218 transcription factor', to specific types of receptor such as 'PC00001 AMPA receptor' and 'PC00030 NMDA receptor'. In order to profile protein classes with human diseases, the PANTHER protein classification system and the corresponding annotation data were downloaded from the PANTHER database on 17/02/2016. An ontology, refereed as PCO (Protein class ontology), was created and subsequently used in disease enrichment analysis. The

annotation data, containing 8852 unique human genes, were aggregated to 97 level 3 PO terms, which were tested with *topOnto* and *HDGDB*. The *elim* topology method was used to estimate the enrichment p-value. The summarized results are shown in fig. A3. The $P_{D,G}$ value for each proteinclass-disease pair was calculated and those with $P_{D,G} > 50\%$ are shown in table 4.3.

The enrichment results again make biological sense and are supported be the literature. 'PC00197 receptor' represents receptors which are a molecules within a cell or on the cell surface characterized by selective binding of a specific substance and a specific physiologic effect that accompanies the binding. As shown in fig. A3f, 'DOID:1561 cognitive disorder' was significantly enriched in 'G-protein coupled receptor' (p=$5.6e-07$), which has been long considered responsible for major cognitive disorder such as schizophrenia and moon disorder [338]. 'PC00084 cytokine receptor' were closely related to 'DOID:2914 immune system disease' (p=$1e-30$), together with 'DOID:934 viral infectious disease' (p=$1.5e-23$) and 'DOID:104 bacterial infectious disease' (p=$4.4e-22$) [339, 340]. Homeostasis of hormone systems is essential for human health, and aberrant regulation of hormone signaling has been associated with many diseases, including cancer, diabetes, and diseases of inflammation. Hormone signaling pathways have become a major interest for pharmacological intervention in many health abnormalities [341]. The hormones androgen and estrogen have long been known to play key roles in development, growth, and homeostasis of reproductive systems and their dysregulation is the major cause of prostate cancer and breast cancer [342, 343]. Besides the two classical steroid hormone receptors, the androgen and estrogen receptors, orphan nuclear receptors are also known to play important roles in tumorigenesis [344] which is consistent with the enrichment of 'DOID:162 cancer' in 'PC00169 nuclear hormone receptor' (p=$e2.4e-07$).

'PC00220 transferase' (fig. A3g) represents any one of a class of enzymes that enact the transfer of specific functional groups (e.g. a methyl or glycosyl group) from one molecule (called the donor) to another (called the acceptor). DNA methylation is an epigenetic modification critical to normal genome regulation and development in which the vitamin folate (or its synthetic form folic acid) is a key source of the one carbon group used to methylate DNA [345]. This explains the result that 'DOID:0080074 neural tube defect' (p=$e4.3e-05$), which is mainly caused by folic acid deficiency [317], was enriched in 'PC00155 methyltransferase'.

'PC00085 cytoskeletal protein' (fig. A3b) represents the major constituents of the cytoskeleton found in the cytoplasm of eukaryotic cells. They form a flexible frame-

work for the cell, provide attachment points for organelles and formed bodies, making communication between parts of the cell possible. 'PC00157 microtubule family cytoskeletal protein' represent proteins that are either microtubules or bind to microtubules to form the cytoskeleton of the cell. It is enriched by 'DOID:0060340 ciliopathy' (p=$4.6e-08$)' which represent a group of disorders associated with either abnormal formation or function of cilia [203]. Cilia are microtubule-based, hair-like cytoplasmic extensions with motile and a range of sensory functions, which are critical for developmental and physiological functions. The microtubule family cytoskeletal proteins are essentially involved in ciliogenesis. In addition, 'DOID:2490 congenital nervous system abnormality' (p=$3.3e-07$), which has recently been linked to cilia defects [310, 311] was also enriched. In the same figure, a host of different immune system diseases were enriched in 'PC00090 defense_immunity protein'. In particular, 'DOID:2914 immune system disease' (p=$6.4e-09$),'DOID:104 bacterial infectious disease' (p=$8.4-09$) and 'DOID:934 viral infectious disease' (p=$1.2e-07$) were enriched in 'PC00149 major histocompatibility complex antigen', a set of cell surface proteins essential for the acquired immune system to recognize foreign molecules in vertebrates, which in turn determines histocompatibility. The main function of the MHC complex is to bind to peptide fragments derived from pathogens and display them on the cell surface for recognition by the appropriate T-cells [346], thus closely related to immune/infectious diseases.

In the protein class 'PC00207 signaling molecule' (fig. A3f), it is found that a wide range of diseases including 'DOID:2914 immune system disease' (p=$1e-30$), 'DOID:15 reproductive system disease' ($2.3e-14$) and 'DOID:28 endocrine system disease' (p=$6e-12$) were enriched in 'PC00083 cytokine'. Cytokines usually act through receptors, and are especially important in the immune system [347]. They are also involved in several developmental processes during embryogenesis [348,349]. 'DOID:162 cancer' ($e4.5e-11$) and 'DOID:0060072 benign neoplasm' (p=$2.2e-10$) were the top two enriched disease found in 'PC00112 growth factor', while for 'PC00179 peptide hormone', closely related diseases such as 'DOID:15 reproductive system disease' (p=$3.9e-14$) and 'DOID:28 endocrine system disease' (p=$5.3e-14$) were highly significantly enriched.

The disease profile of protein classes presents a broad overview of different protein classes and their involvement in the context of human disease. All the genes in the PCO annotation were annotated with at least one human disease from *HDGDB*. No correlation was observed between the size of a protein class and the number of

enriched diseases found within that class (fig. 4.7). Similar to the pathway result, most of the protein classes involved in cell signaling, were enriched with a large number of diseases. For example, ‘PC00083 cytokine’ (33 diseases) and ‘PC00169 nuclear hormone receptor’ (29).

Figure 4.6: The distribution of disease gene and non disease gene in pathways. The pathway is indexed by its unique id for visibility. A surprisingly high number of genes (average 81% across all pathways) annotated with pathways are disease genes that were linked to at least one HDO term.

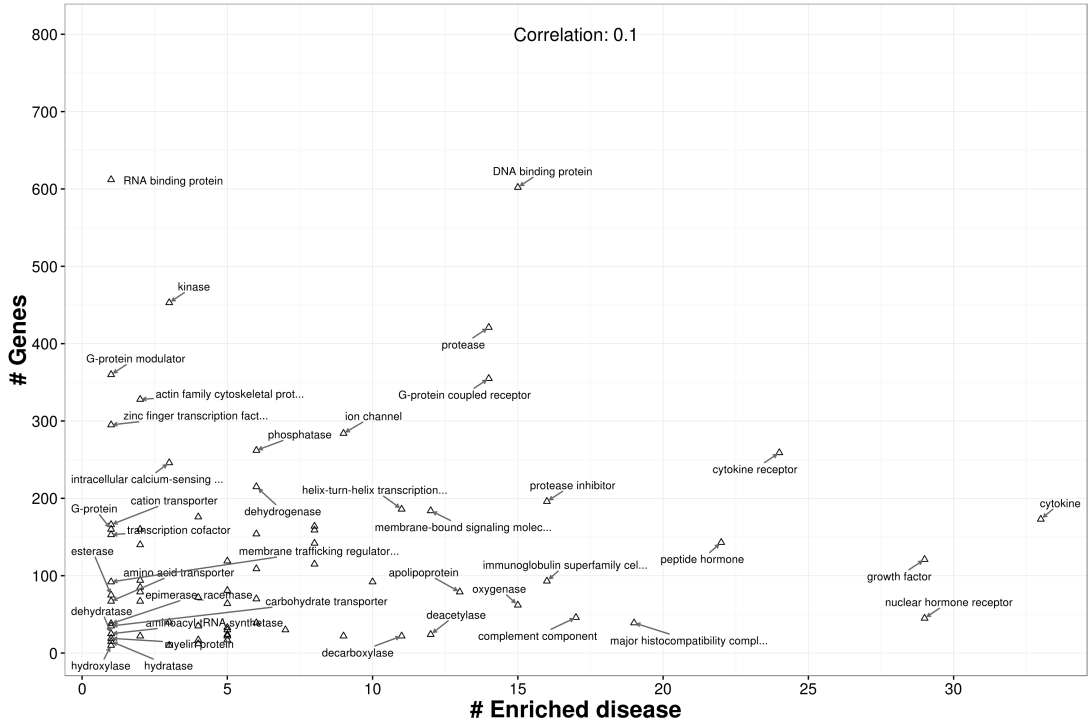


Figure 4.7: Protein class size and the number of enriched diseases. No correlation was observed between the number of protein in a protein class and the number of enriched disease found within that class. The most disease dense protein class was ‘cytokine’ where 33 disease were enriched.

Group	Term.ID	Term.DEF	$N_{D,G}$	N_G	$P_{D,G}$
protein class	DOID:2914	immune system disease	2	3	66.7%
protein class	DOID:934	viral infectious disease	3	3	100.0%
calcium-binding protein	DOID:1287	cardiovascular system disease	2	2	100.0%
calcium-binding protein	DOID:11612	polycystic ovary syndrome	2	2	100.0%
cell adhesion molecule	DOID:16	integumentary system disease	2	2	100.0%
cell junction protein	DOID:16	integumentary system disease	2	2	100.0%
defense_immunity protein	DOID:934	viral infectious disease	2	3	66.7%
defense_immunity protein	DOID:77	gastrointestinal system disease	2	3	66.7%
defense_immunity protein	DOID:2914	immune system disease	2	3	66.7%
defense_immunity protein	DOID:16	integumentary system disease	2	3	66.7%
defense_immunity protein	DOID:104	bacterial infectious disease	2	3	66.7%
defense_immunity protein	DOID:18	urinary system disease	3	3	100.0%
defense_immunity protein	DOID:17	musculoskeletal system disease	3	3	100.0%
defense_immunity protein	DOID:1579	respiratory system disease	3	3	100.0%
extracellular matrix protein	DOID:18	urinary system disease	2	3	66.7%
extracellular matrix protein	DOID:17	musculoskeletal system disease	2	3	66.7%
extracellular matrix protein	DOID:0060072	benign neoplasm	2	3	66.7%
extracellular matrix protein	DOID:0050769	N syndrome	2	3	66.7%
lyase	DOID:655	inherited metabolic disorder	3	4	75.0%
oxidoreductase	DOID:655	inherited metabolic disorder	4	6	66.7%
receptor	DOID:1579	respiratory system disease	3	4	75.0%
receptor	DOID:0060072	benign neoplasm	3	4	75.0%
receptor	DOID:28	endocrine system disease	3	4	75.0%
receptor	DOID:1287	cardiovascular system disease	4	4	100.0%
signaling molecule	DOID:1287	cardiovascular system disease	3	4	75.0%
signaling molecule	DOID:11612	polycystic ovary syndrome	3	4	75.0%
signaling molecule	DOID:934	viral infectious disease	3	4	75.0%
signaling molecule	DOID:16	integumentary system disease	3	4	75.0%
signaling molecule	DOID:17	musculoskeletal system disease	3	4	75.0%
signaling molecule	DOID:1579	respiratory system disease	3	4	75.0%
signaling molecule	DOID:0060072	benign neoplasm	3	4	75.0%
signaling molecule	DOID:15	reproductive system disease	3	4	75.0%
signaling molecule	DOID:28	endocrine system disease	3	4	75.0%
signaling molecule	DOID:18	urinary system disease	4	4	100.0%
signaling molecule	DOID:2914	immune system disease	4	4	100.0%
signaling molecule	DOID:162	cancer	4	4	100.0%
transcription factor	DOID:162	cancer	4	6	66.7%
transferase	DOID:655	inherited metabolic disorder	4	5	80.0%
transporter	DOID:655	inherited metabolic disorder	3	5	60.0%

Table 4.3: Protein class-disease pair with $P_{D,G} > 50\%$.

4.2 Profiling human disease with gene sets

In the above sections, I explored some of the well-established gene sets (created from ontologies) from a disease perspective and tried to answer biological questions like 'what are the most relevant diseases of a particular gene set, for example, for all the protein kinase genes?'. Similarly, using the same method and data, I explore here the molecular basis of human diseases, trying to ask questions like 'what are the most relevant pathways/protein classes that are known to be involved in a disease, e.g. schizophrenia?'.

HDGDB contains gene annotations for 3140 human diseases, out of which 1310 have more than 10 gene annotations (the remaining 1830 diseases were removed because the number of annotations was too small to yield any statistically significant results). Term enrichment analysis was performed on each of the 1310 human diseases against 8 ontologies including 1) RPO, 2) PCO, 3) CO, 4) HPO and the three ontologies from the Gene Ontology namely GOBP (Biological Process), GOMF (Molecular Function) and GOCC (Cellular Component) and HDO itself. The annotation of the ontologies was taken from the corresponding websites on 20^{th} July 2016, except for the HPO annotation, which was generated using OntoSuite-Miner using human phenotype ontology, just like *HDGDB* with HDO. *topOnto* and *HDGDB* were used with the *elim* topology method for analysis. Furthermore, in order to explore diseases that share similar underlying molecular mechanisms, the 1310 diseases were then tested against HDO itself.

By including seven ontologies and HDO itself, a 'disease environment' was created for each of the 1310 human diseases defined by HDO. Two types of node exist in a disease environment, ontology nodes representing the ontology terms and gene nodes represent significant genes. An ontology node is connect to a gene node via edges if the gene is annotated to that ontology term. Each ontology in the disease environment inspects the disease with respect to a different biological theme. For example, for a given disease, the enriched CO terms indicate whether the known disease genes are densely located in a particular chromosome region, while enriched HPO terms reflect the phenotypic abnormalities observed in that disease. A comprehensive over view with possible new insights into each of the 1310 human disease is archived when integrating heterogeneous ontologies into the same environment. In the following section, I will present the disease environment for two diseases, 'DOID:1612 breast cancer' and 'DOID:5419 schizophrenia'. The results of the other disease are available as sup-

plementary information (attached disk) but not presented in the thesis itself.

Note that due to the complicity of the disease environment and it's visibility as a graph in the thesis, only the top 5 enriched terms in each ontology were included. The result of the enriched term were ignored, thus may result in the lost of possible interesting information. More detail analysis could be carried out to use all enriched terms but has not been done in this thesis.

4.2.1 The Disease environment of breast cancer

Breast cancer is the most common invasive cancer in women, affecting about 12% of women worldwide [350] and comprising 16% of all female cancers [351]. About 5-10% of breast cancers are thought to be hereditary, caused by abnormal genes passed from parent to child, including *BRCA1* and *BRCA2* among others. Hereditary breast cancers tend to develop earlier in life than noninherited cases. Some genes such as *TP53, CDH1, PTEN*, *BRCA1* and *BRCA2* are associated with a high risk of developing breast cancer, thus described as 'high penetrance' genes. These genes are usually involved directly in fixing damaged DNA, which helps to maintain the stability of a cell's genetic information. They are described as tumor suppressors because they help keep cells from growing and dividing too fast or in an uncontrolled way. Mutations in these genes usually impair DNA repair, allowing potentially damaging mutations to persist in DNA. As these defects accumulate, they can trigger cells to grow and divide without a control in order to form a tumor. Other genes that have been studied as possible risk factors for breast cancer, either provide instructions for making proteins that interact with the 'high penetrance' genes [352–355], or play a role through other pathways that control the growth, division (proliferation) or apoptosis of cells, or involved in the repair of danged DNA [356]. The combined influence of variations in these genes may significantly impact the risk of developing breast cancer.

In *HDGDB*, 3222 unique genes were annotated to the HDO term 'DOID:1612 breast cancer' and scored based on 26674 pieces of evidence from the three data sources, OMIM, GeneRIF and Ensembl Variation. The top 20 scored GDAs are listed in table 4.4. A series of well known tumor suppressors including *TP53*, *BRCA1-2* and *CHK2* were found amongs the top scoring genes. The *ESR1* gene, encodes estrogen receptor (ER), which is the primary therapeutic target in breast cancer and is expressed in 70% of cases [357]. Mutations in *CDH1* often lead to hereditary diffuse gastric cancer, which is often considered a significant risk factor for breast cancer. *CDH1* dysfunction

may result in a loss of E-cadherin, which may allow breast cells to grow and divide unchecked, leading to a cancerous tumor. Furthermore, since E-cadherin helps neighboring cells stick to one another (cell adhesion) to form organized tissues, the resulting loss of E-cadherin may also make it easier for cancer cells to detach from a primary tumor and spread (metastasize) to other parts of the body [354]. These highly scored breast cancer genes were well studied genes that are closely involved in breast cancer disease mechanism. Together with the lower scoring breast cancer genes, ontology enrichment analysis were performed with 7 ontologies and HDO itself and a breast cancer disease environment was created and is shown in fig. 4.8.

It has been suggested that loci associated with the risk of breast cancer may be distributed non-randomly on chromosomes [287]. However, many previous studies which investigated the association between single nucleotide polymorphisms (SNP) of candidate genes and breast cancer risk yielded inconsistent results, which could be partly due to insufficient power (sample size). It is observed that 17q1 (p=$9.2 - e4$), 20q1 (p=$9.5e - 3$), 1q3 (p=$1.2 - e2$) and 8q2 (p=$1.3 - e2$) chromosome sub bands were enriched in the breast cancer environment, out of which only 8q2 was supported by previous evidence [287]. There are 39, 62 and 43 breast cancer genes located on 17q1, 20q1 and 1q3 respectively, almost one-third of the genes located on 17q1 were annotated to breast cancer. These regions were not identified as breast cancer susceptibility region previously probably due to the different methodology and data used to determine the region's significance. I am using all of the breast cancer genes found in *HDGDB* which also includes those 'low-penetrance' genes and possibly genes that were only associated with breast cancer recently. Thus, the above three chromosome sub bands could be potentially interesting in breast cancer susceptibility and worth further investigation.

'R-HSA-114604 GPVI-mediated activation cascade' (p=$2.9e - 12$), which leads to downstream platelet activation with its downstream pathway 'R-HSA-76005 Response to elevated platelet cytosolic Ca2+' (p=$4.4e - 10$), has been identified as the most enriched pathway for breast cancer. Previous studies have shown that an increased circulating platelet count (thrombocytosis) is associated with a poor breast cancer prognosis, suggesting a potential direct role for platelets in the pathogenesis of breast cancer [358, 359]. In fact, direct evidence exists, supporting the close involvement of platelet in 6 out of the 10 hallmarks [305] that defines cancer pathogenesis [360–363], which is also consistent with the RPO disease profile of 'R-HSA-76002 Platelet activation, signaling and aggregation' (fig. A2c) enriched with 'DOID:162

cancer' (p=$1.80e-07$). Thus together with its preponderance in the breast cancer disease environment, the current under-explored anti-platelet therapy could be promising adjunct to existing breast cancer treatments. The presence of 'R-HSA-2219530 Constitutive Signaling by Aberrant PI3K in Cancer' and 'R-HSA-1257604 PIP3 activates AKT signaling' were interesting. The PI3K/AKT signaling is frequently constitutively activated in cancer via gain-of-function mutations in one of the two PI3K subunits, PI3KCA or PIK3R1. Mutations in PIK3CA enable the kinase to achieve an active conformation, producing PIP3 and activating downstream AKT in the absence of growth factors. Aberrant gain of Akt activation underlies the pathophysiological properties of a variety of complex diseases, not only breast cancer but also cancers in general [329, 364].

Different types of cancer are presented in the breast cancer environment because cancer often share similar pathogenesis such as uncontrolled cell growth and division. Thus they are likely to share a large amount of genes. Interestingly, even though breast cancer is rare in males, 'DOID:10283 prostate cancer' (p=$1e-30$) was present in the breast cancer environment. This, beside the general cancer mechanism, is probably because in a small number of men, prostate cancer is linked to alterations in the BRCA1 or, more often, the BRCA2 gene [365, 366]. In fact, a recent study found that a family history of prostate cancer was associated with a modest increase in breast cancer risk, which suggests that prostate cancer diagnosed among first-degree family members increases a woman's risk of developing breast cancer [367]

4.2.2 The Disease environment of schizophrenia

Schizophrenia is a disease characterized by a disintegration of the processes of thinking and of emotional responsiveness. It ranks among the top 20 causes of disability worldwide and caused more loss of life than several cancers and other physical illnesses [368]. A number of theories have been developed, attempting to explain how changes in brain function can contribute to symptoms of the disease [369–372] but the underlying mechanisms of schizophrenia are complex and is not entirely clear yet. In an attempt to gain a better understanding of schizophrenia, ontology enrichment analysis was performed using schizophrenia genes against 7 ontologies and HDO itself and a schizophrenia disease environment was created, representing the current knowledge of the disease and possible insights. The schizophrenia disease environment is shown in fig. 4.8.

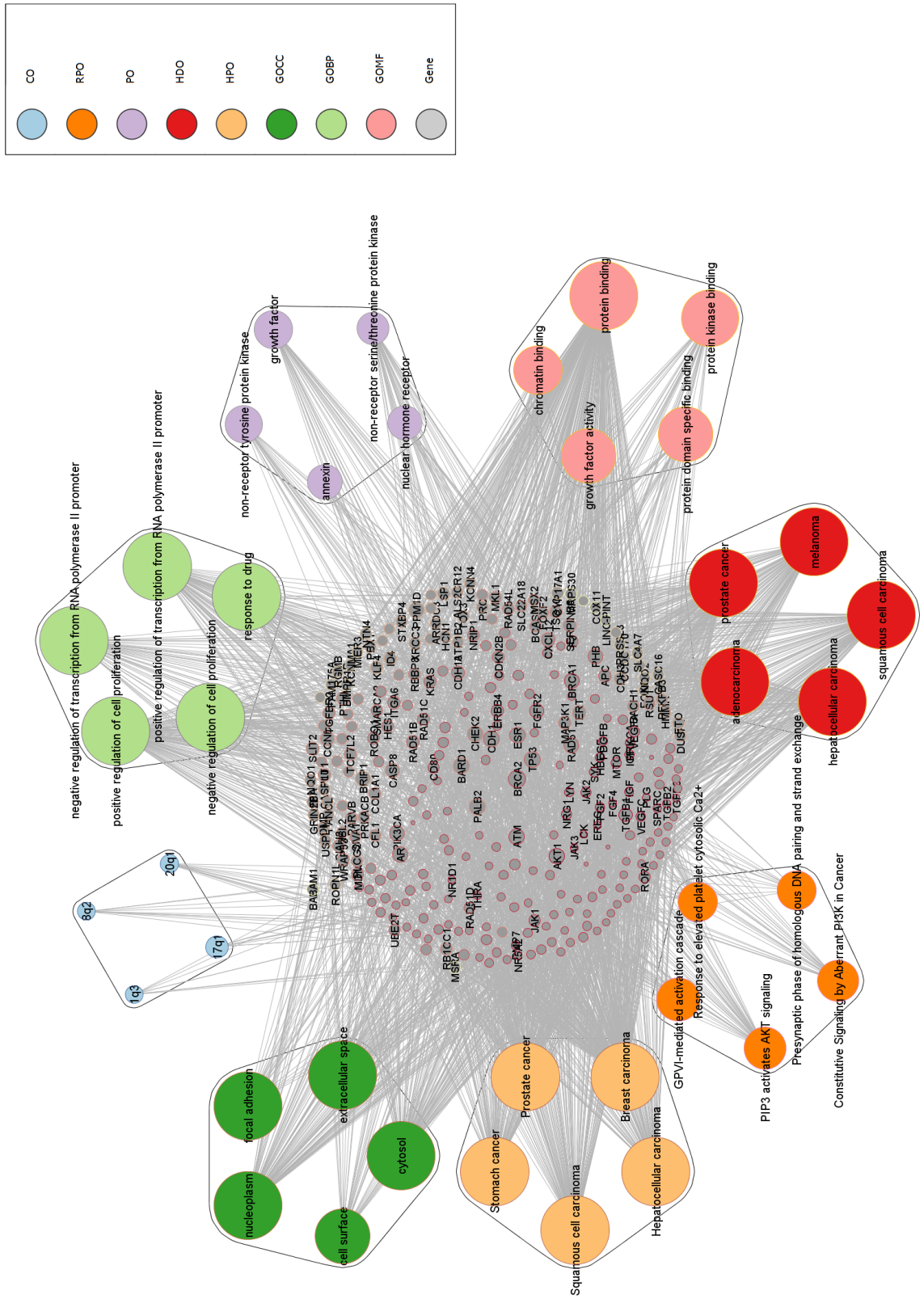


Figure 4.8: Disease environment for breast cancer. An edge is placed between a gene and a term if the gene is contributing to the enrichment of the term. The top 5 enriched ontology terms were grouped and sized base on their significant value (p-value). Gene nodes were sized base on their GDA scores with breast cancer and border colored based on the number of ontologies they are involved. For visibility, low scored gene nodes (< 0.3) which only contributing to less than 6 ontologies were removed from the graph.

	DOID	Disease	Gene	Score	Score.o	Score.g	Score.v	#Evidence (o/g/v)
1	DOID:1612	breast cancer	BRCA2	0.84	0.217	0.319	0.259	2/273/2466
2	DOID:1612	breast cancer	ESR1	0.72	0.167	0.318	0.196	1/373/11
3	DOID:1612	breast cancer	TP53	0.72	0.167	0.317	0.167	1/221/1
4	DOID:1612	breast cancer	CDH1	0.71	0.167	0.314	0.188	1/57/3
5	DOID:1612	breast cancer	PALB2	0.71	0.167	0.314	0.196	1/43/11
6	DOID:1612	breast cancer	ATM	0.71	0.167	0.311	0.192	1/46/5
7	DOID:1612	breast cancer	AKT1	0.71	0.167	0.308	0.182	1/86/2
8	DOID:1612	breast cancer	AR	0.71	0.167	0.307	0.188	1/48/3
9	DOID:1612	breast cancer	BARD1	0.70	0.167	0.303	0.182	1/15/2
10	DOID:1612	breast cancer	PPM1D	0.69	0.167	0.299	0.182	1/10/2
11	DOID:1612	breast cancer	PIK3CA	0.66	0.167	0.258	0.188	1/71/3
12	DOID:1612	breast cancer	CHEK2	0.66	0.167	0.258	0.182	1/61/2
13	DOID:1612	breast cancer	RAD51	0.65	0.167	0.249	0.182	1/33/2
14	DOID:1612	breast cancer	BRIP1	0.65	0.167	0.25	0.192	1/11/5
15	DOID:1612	breast cancer	BRCA1	0.64	0.05	0.319	0.259	1/484/1914
16	DOID:1612	breast cancer	PHB	0.64	0.167	0.242	0.188	1/5/3
17	DOID:1612	breast cancer	CASP8	0.59	0.167	0.198	0.182	1/19/2
18	DOID:1612	breast cancer	FGFR2	0.57	0	0.304	0.252	0/44/16
19	DOID:1612	breast cancer	TGFB1	0.51	0	0.312	0.167	0/74/1
20	DOID:1612	breast cancer	CCND1	0.51	0	0.308	0.19	0/68/4

Table 4.4: Top scored gene-disease associations for breast cancer in the Human Disease Gene Database.

In *HDGDB*, 1844 unique genes were annotated to the HDO term 'DOID:5419 schizophrenia' and scored base on 3920 pieces of evidence from the three data sources, OMIM (29), GeneRIF (2441) and Ensembl Variation (1450). The top 20 scored GDAs are listed in table 4.4. *COMT* (Catechol-O-methyltransferase) is a gene that produces an enzyme that helps maintain appropriate levels of certain neurotransmitters such as dopamine and norepinephrine at the prefrontal cortex. Most studies have focused on the effects of a particular common variation in *COMT*, which alters a single protein building block in the enzyme, replacing the amino acid Valine with the amino acid Methionine subsequently affecting the enzyme's ability to break down neurotransmitters in the prefrontal cortex. Since dopamine has long been considered as central to the pathophysiology of schizophrenia [373], *COMT* is a clear functional candidate gene for Schizophrenia and has received a lot of research attention because its involvement in the regulation of dopamine. For the same reason, the *DRD2* and *DRD3* gene are scored highly in the list because they encode the D3 and D2 subtypes out

of the five (D1-D5) subtypes in dopamine receptors. Other top scored genes including *DISC1*,*MTHFR* and *PRODH* are all thought to be associated with susceptibility schizophrenia [374–376].

Interestingly, the two genes *HTR3A* and *HTR3B* were found in the schizophrenia disease environment, contributing to the enrichment of terms across all of the 8 ontologies with relatively low scores. *HTR3A* and *HTR3B*, both scored 0.188, were associated to schizophrenia by *OntoSuite-Miner* based on 4 (2 each) GeneRIFs from the GeneRIF database from 3 recent PubMed entries [377–379]. In humans, the two genes encode two of the five subunits (HTR3A-E) of the 5-hydroxytryptamine receptor 3 (HTR3), a group of G protein-coupled receptors (GPCRs) and ligand-gated ion channels found in the central and peripheral nervous system that have been suggested to play a role in the pathophysiology and treatment of schizophrenia [380, 381]. Variants of HTR3 genes have been associated with treatment outcomes of antipsychotics in schizophrenia [381]. Genetic association studies have shown that the chromosome 11q22-24 region, which is enriched ($p=2e-04$) in the schizophrenia environment, is a susceptibility locus for schizophrenia [382]. *HTR3A* and *HTR3B* are the two out of the five HTR3 genes that located within the region of chromosome 11q23.1, which makes them most interesting for schizophrenia studies. A case study with 140 schizophrenia patients taking clozapine (a medication that treats schizophrenia) for 6 months revealed significant allelic association of clozapine response with three variants in the *HTR3A* receptor and suggested that variants in the *HTR3A* receptor gene can play a role in the treatment outcome of clozapine in schizophrenia patients that are refractory or intolerant of atypical antipsychotic therapy [378]. Another study with 222 Korean schizophrenia patients [377] found an association of the *HTR3B* with poor concentration in schizophrenia patients, suggested that the *HTR3B* may be involved in the attention problem of schizophrenia in the Korean population. These recent studies consider a small number of subjects, thus the findings are preliminary and need to be further validated. The high involvement of the *HTR3A* and *HTR3B* in the schizophrenia environment not only supports the importance of these two genes in schizophrenia, but also suggests that low score genes might be important in the disease especially when the low score is due to the fact that they come from very recent findings. Such property of low scored genes in *HDGDB* is likely to be true for all diseases explored in this way.

The top enriched pathway in the schizophrenia disease environment was 'R-HSA-438066 Unblocking of NMDA receptor, glutamate binding and activation' ($p=8.7e-$

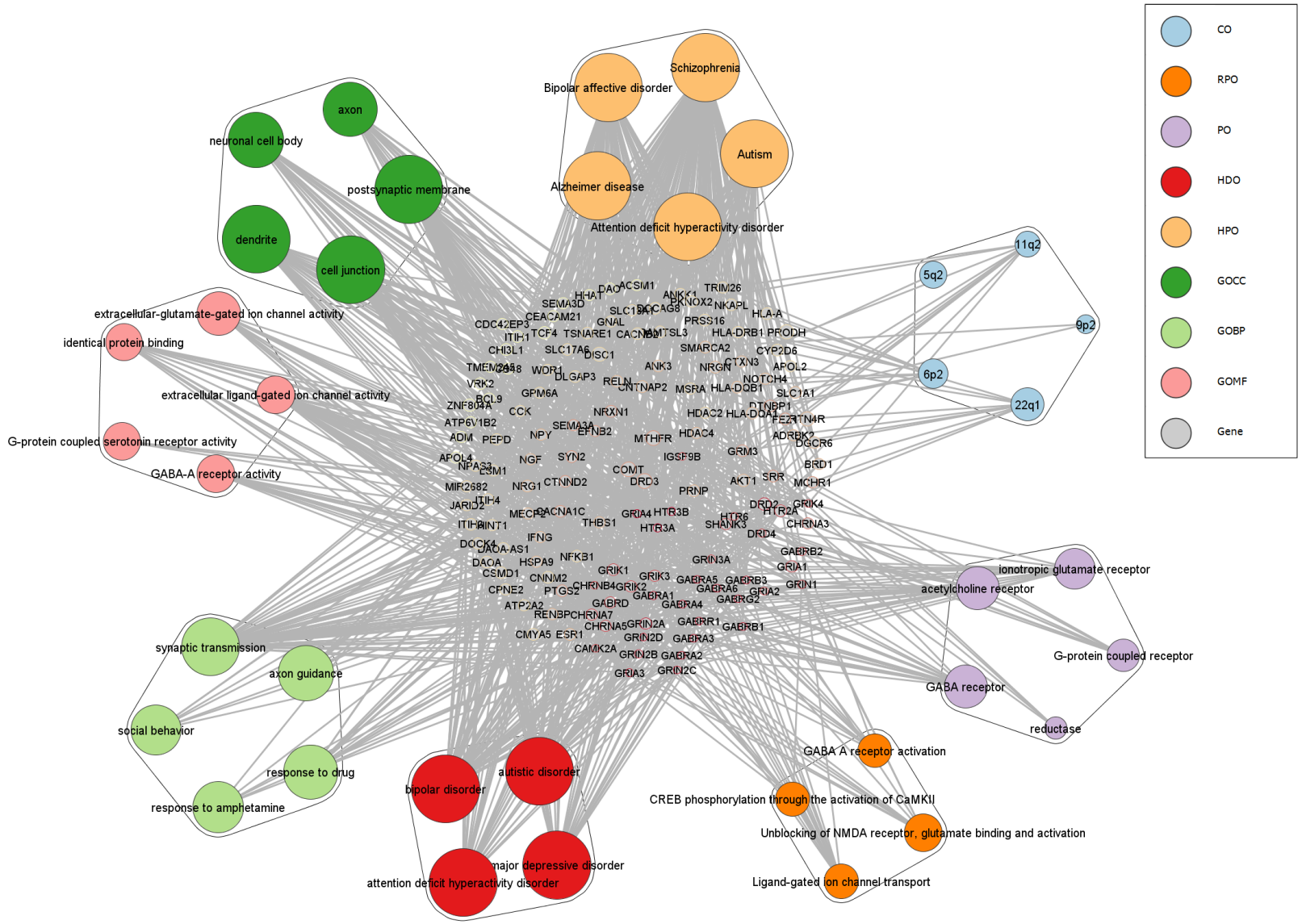


Figure 4.9: Disease environment for schizophrenia. Nodes represent ontology terms or genes. An edge is placed between a gene node and an ontology node if the gene is contributing to the enrichment of that term. The top 5 enriched ontology terms were grouped and sized base on their significant value (p-value). Gene nodes were sized base on their GDA scores with schizophrenia and border colored based on the number of ontologies they are involved. For visibility, low scored gene nodes (score< 0.3) which only contributing to less than 6 ontologies were removed from the graph.

	DOID	Disease	Gene	Score	Score.o	Score.g	Score.v	#Evidence (o/g/v)
1	DOID:5419	schizophrenia	COMT	0.65	0.167	0.25	0.182	1/98/2
2	DOID:5419	schizophrenia	DRD3	0.65	0.167	0.248	0.182	1/24/2
3	DOID:5419	schizophrenia	DISC1	0.60	0.167	0.199	0.182	1/68/2
4	DOID:5419	schizophrenia	MTHFR	0.60	0.167	0.199	0.167	1/29/1
5	DOID:5419	schizophrenia	PRODH	0.59	0.167	0.194	0.197	1/7/15
6	DOID:5419	schizophrenia	RTN4R	0.59	0.167	0.194	0.19	1/6/4
7	DOID:5419	schizophrenia	DRD2	0.45	0	0.249	0.182	0/57/2
8	DOID:5419	schizophrenia	ZNF804A	0.45	0	0.249	0.19	0/35/4
9	DOID:5419	schizophrenia	AKT1	0.45	0.167	0.247	0	1/16/0
10	DOID:5419	schizophrenia	TCF4	0.45	0	0.246	0.195	0/10/8
11	DOID:5419	schizophrenia	NRG1	0.40	0.167	0.199	0	1/75/0
12	DOID:5419	schizophrenia	DTNBP1	0.40	0.167	0.199	0	1/48/0
13	DOID:5419	schizophrenia	DAOA	0.40	0.167	0.199	0	1/36/0
14	DOID:5419	schizophrenia	HTR2A	0.40	0.167	0.199	0	1/28/0
15	DOID:5419	schizophrenia	CACNA1C	0.40	0	0.197	0.194	0/12/6
16	DOID:5419	schizophrenia	GRM3	0.40	0	0.198	0.167	0/18/1
17	DOID:5419	schizophrenia	RELN	0.40	0	0.198	0.167	0/18/1
18	DOID:5419	schizophrenia	DAO	0.40	0.167	0.197	0	1/14/0
19	DOID:5419	schizophrenia	DGCR6	0.40	0	0.182	0.197	0/2/15
20	DOID:5419	schizophrenia	ANK3	0.39	0	0.195	0.192	0/8/5

Table 4.5: Top scored gene-disease associations for schizophrenia in the Human Gene Disease Database.

09). The neurotransmitter glutamate and the reduced function of the NMDA glutamate receptor has been hypothesized in the pathophysiology of schizophrenia, which was largely suggested by abnormally low levels of glutamate receptors found in postmortem brains of people previously diagnosed with schizophrenia [371] and the discovery that glutamate blocking drugs can mimic the symptoms and cognitive problems associated with schizophrenia [383]. The enriched pathway 'R-HSA-977441 GABA A receptor activation' (p=$4.7e-07$) supports an alternative schizophrenia pathophysiology hypothesis which was recently proposed, that of dysfunction of interneurons (GABAergic) in the brain. The GABAergic system is principally involved in the balance of excitation and inhibition in the brain. GABA is synthesized in 20-30% of all central nervous system (CNS) neurons and is the primary transmitter at 25-50% of synapses in mammalian brain. Its ubiquity means that almost all neurons express GABA receptors, thus it is expected that most brain functions involve GABAergic transmission [384]. GABA has three known classes of receptors: GABA-A, GABA-

B and GABA-C receptors. GABA-A and GABAC receptors are ligand gated ion channels (ionotropic), while the GABAB receptors are G-protein coupled receptors (metabotropic). In particular, GABA-A receptors are the most complex of the three, and play a critical role in brain development, facilitating neuronal migration and regulating differentiation and synaptogenesis, and mediating both fast and tonic GABAergic inhibition. Changes in the function and expression of these receptors have been strongly implicated in schizophrenia [385–387] and their therapeutic potential has been explored [388]. This was also captured by the schizophrenia environment by the co-enrichment of terms from multiple ontologies including 'GO:0004890 GABA-A receptor activity' (p=$9.1e-09$), 'GO:0045211 postsynaptic membrane' (p=$1e-30$), 'GO:0030054 cell junction' (p=$1.0e-25$), 'GO:0007268 synaptic transmission' (p=$1.5e-19$), 'GO:0042493 response to drug' (p=$1.5e-17$), 'R-HSA-975298 Ligand-gated ion channel transport' (p=$2.1e-07$) and 'PC00023 GABA receptor' (p=$4.6e-11$). The precise role GABA plays in the pathogenesis of schizophrenia is not entirely clear. However, GABA appears to have an effect on regulation of dopamine levels in the brain, which is an inhibitory neurotransmitter involved in the pathology of schizophrenia that is also enriched with 15 terms including 'R-HSA-379397 Enzymatic degradation of dopamine by COMT' (p=$1.9e-2$) and 'GO:0042053 regulation of dopamine metabolic process' (p=$7.0e-07$) (not shown in the Figure). There is a growing body of research suggesting that GABA-dopamine interaction is responsible for some symptoms of schizophrenia [389].

In terms of enriched chromosome sub bands, the non-random distribution of polymorphic loci associated with breast cancer [287] suggested that human chromosomes are heterogeneous in structure and function. Schizophrenia is a common and serious psychiatric illness with strong evidence for genetic causation, thus chromosomal abnormalities associated with schizophrenia may help to understand the genetic complexity of the illness. It is observed in the schizophrenia disease environment that 22q1 chromosome sub band was the top enriched chromosome sub band for schizophrenia. This is probably due to 22q11.2 deletion syndrome (22qDS), which is the only genetic form of schizophrenia that is recurrent, clinically recognizable, and has confirmatory genetic testing available [290]. Even though no single 22q11.2 deletion region is necessary and sufficient for expressing the major features of 22qDS [390], a typical 3-Mb hemizygous 22q11.2 deletion is most commonly associated with schizophrenia [391]. 6p2 chromosome sub band was identified as the second enriched chromosome sub band in the schizophrenia disease environment. Similar results have been found in [288]

where 39 loci associated with schizophrenia risk were tested and the result revealed that chromosome segments 6p21.1-p22.3 bear a significantly higher number of susceptible loci. Furthermore, in 2011, a meta-analysis of genome-wide association studies discovered that 129 out of 136 single-nucleotide polymorphisms (SNP) significantly associated with schizophrenia were located in the major histocompatibility complex (MHC) region which occurs on chromosome 6 from 6p22.1 to 6p21.3 [392]. Such association of schizophrenia and the MHC complex suggests a connection of the disease to the human immune system, which has been proposed recently as a novel hypothesis of schizophrenia pathophysiology [393, 394]. Chromosome sub band 5q2, 11q2 and 9p2 were enriched, where 22, 46 and 19 genes were associated with schizophrenia but there is no evidence found in the literature that statistically tests these regions for schizophrenia susceptibility. Such results suggest possible novel chromosome sub bands for schizophrenia susceptibility but further investigations is needed , for example, using genetic screening methods to find and compare mutations in the genes found in these sub bands between people with and without schizophrenia.

The enrichment of HDO terms highlighted a similar genetic signature between these diseases and schizophrenia. 'DOID:3312 bipolar disorder', 'DOID:1094 attention deficit hyperactivity disorder', 'DOID:12849 autistic disorder' and 'DOID:1470 major depressive disorder' are all mental disorders with symptoms similar to those seen in schizophrenia. Roughly 50% of the genes annotated to these diseases were also associated with schizophrenia, indicating a large shared underlying molecular mechanisms for these diseases. Such overlapping also result in the overlap of shared/similar clinical features/symptoms that make diagnostically differentiating between these diseases difficult. For example, childhood-onset schizophrenia (COS), considered a rare and severe form of schizophrenia, frequently presents with premorbid developmental abnormalities that exhibit clinical features includes deficits in communication, social relatedness, and motor development, similar to those seen in autism spectrum disorders. 'Schizoaffective disorder', which is an intermediate diagnosis for patient has features of both schizophrenia and a mood disorder, either bipolar disorder or depression, but does not strictly meet diagnostic criteria for either alone.

4.2.3 Exploration of connection between human disease

In order to explore the inter connection between diseases, I defined $N_{A,B}$ to be the number of times disease A and disease B found enriched at the same time among the

top 5 diseases in a gene class. I calculated N for each enriched disease pair in the 277 gene classes tested, a summary of results ($N > 5$) where is shown in fig. 4.10. Interestingly, it is observed that some disease groups were enriched together more often than others. 'immune system disease' was found to coexist with a range of diseases including 'gastrointestinal system disease'($N = 19$) and 'viral infectious disease'($N = 18$), reflecting the importance of the immune system. 'cancer' was enriched together with 'benign neoplasm'(18), 'disease of cellular proliferation'(8) and 'pre-malignant neoplasm'(8). These observations identify a group of closely related diseases, some of which are well known to be connected while others are clinically distinct diseases. Coexistence suggests that diseases may have a similar or share common underlying molecular mechanisms.

Even through the annotation of the 277 gene classes was created separately by the corresponding ontology consortium, the disease profile shows a consistent result for similar gene classes across different ontologies. This overlap is mostly between PCO and RPO. For instance, in the gene class group level (level 2 ontology terms), recall that I defined $P_{D,G}$ which is an indication of a disease D being consistently enriched across group G, there are $P_{cancer,signalingmolecule} = 100\%$ in PCO vs $P_{cancer,SignalTransduction} = 83\%$ in RPO. In the gene class level(level 3 ontology term), 'neural tube defect' was enriched in PCO 'methyltransferase' and in RPO including 'Metabolism of vitamins and cofactors' and 'Biological oxidations'. Such consistency increases confidence in disease mechanisms, as well as the utility of the *HDGDB* disease annotation database.

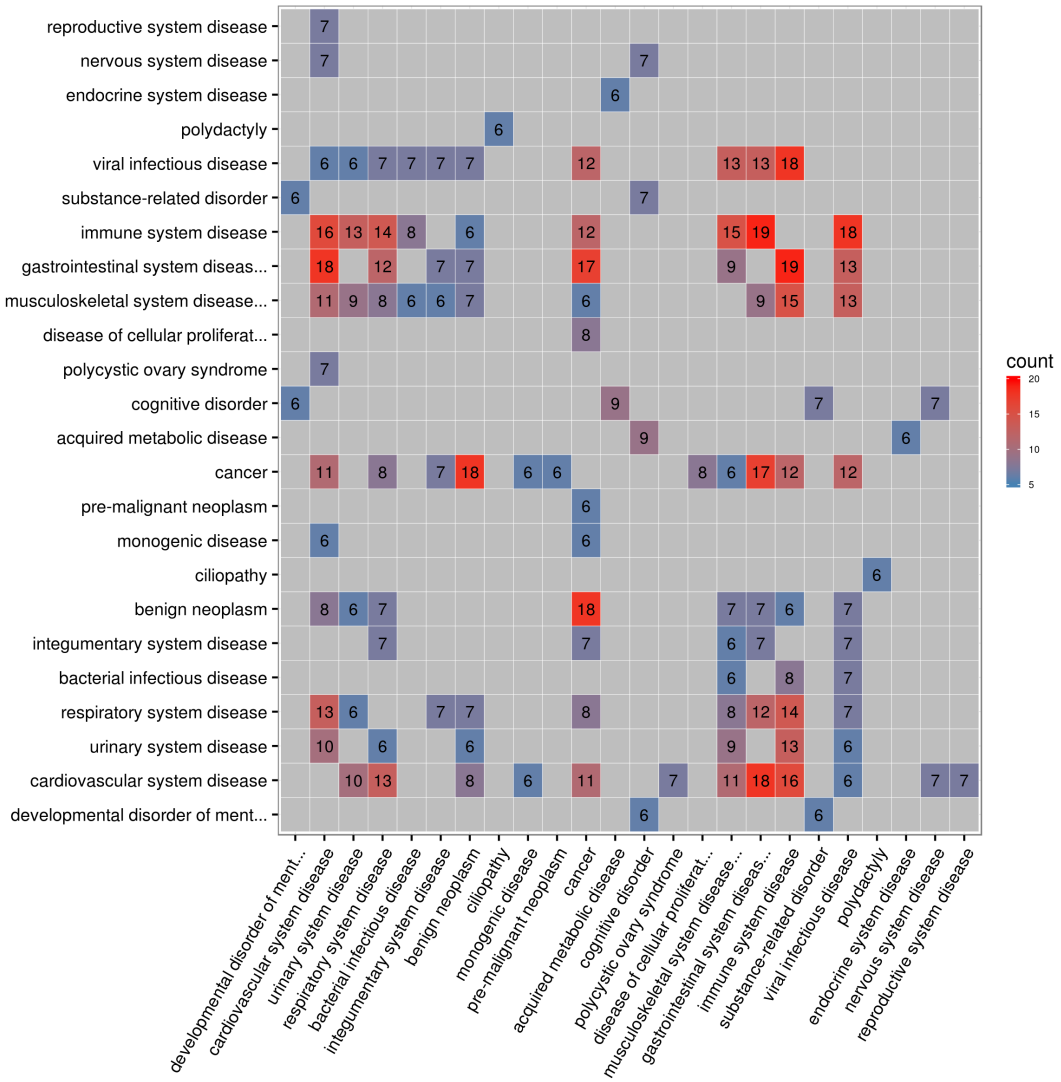


Figure 4.10: The number of co-existing disease pairs among the top enriched diseases between the 277 gene classes.

4.3 Conclusions

In the above section, I used disease enrichment analysis with the created gene disease annotation, *HDGDB*, to profile 277 gene classes, constructed from level 3 ontology terms from the three ontologies including CO, PCO and RPO. The results, firstly, provided an overview of the importance/involvement of each gene class in the context of human disease, in which most of them make biological sense and were supported in the literature; and secondly, demonstrated that the integration of disease data and other divergent information can recover well known disease knowledge as well as give potentially new insight into some of the less well known diseases.

Similarly, I explored the molecular basis of human diseases, created a 'disease environment', with the top 5 enriched terms in each ontology, for each of the 1310 human disease with 8 ontologies including 1) RPO, 2) PCO, 3) CO, 4) HPO, 5) HDO and the three ontologies from the Gene Ontology namely GOBP (Biological Process), GOMF (Molecular Function) and GOCC (Cellular Component). In particular, I presented the disease environment for two diseases, 'breast cancer' and 'schizophrenia'. A lot of important aspects of the two disease were inspected, discussed and supported by literature. The result suggests that the integration of multiple biological ontologies is potentially more powerful than using single one and the 'disease environment' is an interesting analytical approach to systematically gather knowledge for disease with the potential of gaining new insight into the disease mechanisms by transferring knowledge from other close related diseases.

Individual gene level statistics were not discussed in this section because the level 3 ontology terms are aggregated from their child terms. For example, for disease annotation, the term 'cancer' contains all the genes that were annotated to different types of cancer, including those genes that only affect specific types of cancer. Thus, 'cancer' is a representation of all known cancer related genes in the *HDGDB* data; while for the protein class 'cytokine', it contains annotation from sub protein classes including 'TGF-beta superfamily member', 'chemokine' and 'tumor necrosis factor family member'. The disease profiling is intended to generate an overview of the tested gene class. For the same reason, there is a lot of confirmatory (expected), but not novel, observations in the disease profiles. However, it is expected that more insightful results will emerge when profiling lower level terms, the leaf terms for example, from the ontologies.

Chapter 5

Conclusions

Ontologies are semantic frameworks upon which biological data can be structured and have grown to be one of the great enabling technologies of modern bioinformatics. The transformation of unstructured to structured data using data mapped to ontologies has largely been achieved in a time-consuming manner which relies on human experts. Such manual data curation is accurate but hard to scale and unable to keep pace with the rapid expansion and refinement of both ontologies and the data that we would like to annotate to them. Many bioinformatic tools only support analysis using the Gene Ontology because it is the best annotated and is the most widely used and cited. The delay annotating new ontologies and the lack of support for them in existing analytical tools to aid biological interpretation of data has become a major limitation to their utility and uptake. The work presented in this thesis aimed to develop automatic approaches to facilitate the transformation of unstructured data to unlock the potential of all ontologies, with corresponding bioinformatics tools to support their interpretation. The proposed *OntoSuite* framework, provided not only the transformation of unstructured data to ontology base data using scalable text mining approaches, but also the corresponding bioinformatics tools to support the usage of the resulting ontology based annotation. Human disease ontology was used to defenestrate the functionalities of the framework but the usage of the framework is not limited to any particular ontology. In this concluding chapter, I will discuss briefly the value and limitations of the work presented, together with a speculative discussion on possible improvements. More in depth discussion of limitations can be found in the conclusion section of each chapter.

In chapter 1 on page 1, I introduced the importance and the challenges of transforming unstructured biological data into ontology based annotation and reviewed re-

lated work in the field that contributed to the challenge. The motivation of the work presented in the thesis was summarized.

In chapter 2 on page 33, I proposed the first part of the *OntoSuite* framework named *OntoSuite-Miner*, which uses natural language processing approaches for the integration of biological text corpora onto unifying ontologies that structurally represent biological information across a whole host of domains. Two of the most commonly used nature language processing toolkits in the bio-text mining domain, namely Metamap(*MeM*) and NCBO Annotator(*NcA*) were integrated into *OntoSuite-Miner*, with the possibility of adding extra annotators such as the Concept Mapper. The performance of *OntoSuite-Miner* was evaluated using manually curated GWAS data which result in a better F score than using *MeM* and *NcA* alone, indicating the benefit of the integration. Three publicly available biomedical databases, OMIM, GeneRIF and Ensembl Variation, and created a human gene disease annotation dataset, named *HDGDB*, with the Human disease ontology (HDO). Scores were assigned to each gene disease annotation based on the evidences from the three data sources. All of the supporting evidence is linked to the corresponding gene disease association, allowing the user to trace back to the original source of information and to explore the association in its original context. These aspects are of crucial importance in evaluating the evidence supporting a scientific assertion, in order to determine its relevance in the user context. Due to the fact that the annotation was automatically generated based on natural language processing algorithms, errors are inevitable. Thus, I assessed and evaluated *HDGDB*, and discussed limitations of the approach and the possible future improvements. Since biomedical databases are regularly updated, *OntoSuite-Miner* can be configured to update automatically so that the annotation database can be kept up to date. *OntoSuite-Miner* has been implemented and applied to generated Human Disease Ontology(HDO) annotation via publicly available bio-medical text corpus. However, the usage of *OntoSuite-Miner* is not limited to HDO. The advantage of the framework is that it allows the creation of any ontology based annotation given a proper text corpus. This functionality, to a certain degree, facilitates the usage of some of the less used ontologies and enables the usage of customized ontologies in gene annotation.

In chapter 3 on page 113, I reviewed the different statistical methods used in the enrichment analysis and proposed the second half of the *OntoSuite* framework, *OntoSuite-Analytics*, which integrates a collection of enrichment analysis methods into a unified R package named *topOnto*. The package supports enrichment analysis across multiple ontologies with a set of implemented statistical/topological algorithms, allow-

ing comparison of enrichment result across multiple ontologies. Besides the standard over-represent enrichment analysis methods and GSEA, I also proposed the GSEA-CSW algorithm which takes into account a annotation confident score when calculating the enrichment p-value. The algorithm was particularly designed for gene set enrichment analysis when the back-end annotation data are scored, such as those generated in *HDGDB*. Validation of the algorithm was performed on simulated gene expression data. The result indicated that the GSEA-CSW performs as good as the original GSEA algorithm, and it is able to capture the different of annotation confidence score when they are available. Limitations of *OntoSuite-Analytics* and possible future work was also discussed at the end of chapter 3 on page 113.

In chapter 4 on page 149, I systemically profiled human genes with human diseases using *HDGDB* and *OntoSuite-Miner*. I demonstrated the value of integrating multiple ontologies into enrichment analysis, showing how orthogonal ontologies can improve the interpretation of biological data which would have been missed by using a single ontology.

5.1 Limitations and future work

A few points about the implementation and methodology need further discussion and improvement. The current version of *OntoSuite-Miner* only returns a binary connection between entities in the text corpus. Thus it can not provide information about the level of certainty between the entities. In other words, the same link between gene A and disease B will be constructed for the following three texts: 1) 'Gene A is associated with disease B', 2) 'Gene A might be associated with disease B' and 3) 'Gene A is over expressed in disease B'. Several approaches have been proposed to overcome the problem in the field of relation extraction (RE) including rule-based approaches [133,234], co-occurrence based statistic approaches [154,235,236], machine learning [45, 237–240], NLP-based systems [241, 242]. Integrating one or a few of these approaches to *OntoSuite-Miner* would allow the capture of an extra layer of data and could potentially increase the usefulness of the resulting gene annotation data.

In addition, *OntoSuite-Miner* implements two of the most commonly used nature language processing toolkits in the bio-text mining domain, namely Metamap and NCBO Annotator. A bio-text/ontology mapping is considered more reliable if it is agreed upon by both concept recognizers simultaneously. Under this rationale, it is possible to add more concept recognizer modules into the framework. One possible

alternative is the Concept Mapper [147], which is a tool that was not specifically developed for biomedical term recognition but performs very well in biomedical text mining tasks base on the review by Funk et.al. [144]. With extra concept recognizer modules, *OntoSuite-Miner* could potentially generate more accurate results, which in turn benefits downstream analysis. What's more, since NCBO is the central repository for bioontologies, it would be useful to synchronize ontologies with NCBO and automatically add them into the *OntoSuite* framework. This would greatly increase the number of supported ontologies and ease the burden for users for parsing/loading ontologies into the framework.

A confidence score has been implemented, indicating the level of certainty of each gene disease association(GDA) generated from *OntoSuite-Mner*. The score takes into account the number of sources, the amount of evidence that supports the association and the number of annotators that identified the association. It provides a way to rank/weight the associations based on the evidence and assists in the prioritization and navigation of the GDAs. In the current implementation, the sources are equally weighted, which is not a true reflection of the quality of the source. For example, gene disease associations identified in GeneRIF are likely to be less precise than a human curated sources such as OMIM. However, it is not obvious how to weight theses sources when no training data is available to estimate the weight. Thus, a better estimation of the GDA score can be achieved when such data become available.

What's more, In section 2.2.2, genes were linked to diseases by SNP based on their location on the chromosome. Currently, the closest up/down stream and the overlapping genes of a SNP were linked to the disease/phenotype via the SNP. The reason of including the closest up/down stream blindly is due to the lack of understanding of non-coding genome. As the improvement of screening technology, the functions of non-coding region are being revealed. The gene-SNP mapping could be improved as such data become available. For example, gene can be weighted by different factors regarding the relation of the gene to the SNP. A higher weighted could be assign to the gene-SNP mapping If the SNP is located in the regulatory region of the gene.

There are other aspects of the implementation that could be improved regarding *OntoSuite-Analytics*. First of all, *topOnto* is currently implemented in R, which requires users to have some experience of programming or using the command line. It would be more convenient to provide a user friendly web interface for the user to simply upload their gene list, choose the algorithm and get the results without needing to consider the code.

Secondly, the newly implemented GSEA-CSW method was implemented, and tested only with syntactic data.. This is because evaluating the performance of the algorithm is difficult due the absence of gold standards. In [16], Huang et.al developed a method to evaluate different enrichment methods on a data set generated by randomly shuffling the phenotype labels of an experimental data set. The rationale behind the method is that a gene set deemed significantly enriched by more statistical methods is less likely to be false than a gene set deemed significant by fewer statistic methods. An MC (mutual coverage) score is computed for each statistical method against others representing the degree of mutually identified enriched gene set between them. This methods cannot be directly applied to evaluate GSEA-CSW due to the lack of standard weighted annotation data. GSEA-CSW is a modified version of the original GSEA algorithm, which itself has been carefully tested. The difference between the GSEA and GSEA-CSW is that GSEA-CSW changes the step length based on the annotation score, thus increasing the contribution of highly weighted genes in the enrichment analysis. A scaling factor λ was introduced in the algorithm to control the magnitude of the effect of the annotation score, but the optimized value of λ was not determined. It is still unclear how this value can be estimated due to the complicity of real biological data, thus further work is needed to estimate the λ.

Thirdly, the well recognized multiple hypothesis testing problem [55, 90, 94–97] is not solved in *OntoSuite-Analytics*. This is because the multiple hypothesis correction itself is still an open research question and it has a relatively small impact compared to other factors that affect enrichment analysis, such as the topology algorithm or background annotation data used. The independent test assumption does not fit into the enrichment test where terms are interconnected in an ontology. As a result, the current implementation of *topOnto* returns the raw p-values as a result but provides options for the user to choose predefined multiple hypothesis correction methods based on their needs. If the user aims to find out a set of enriched processes for further experimental validation, then a simple Bonferroni correction could be applied to produce the most conservative result to prevent false positive results being tested further in expensive biological experiments. If the analysis is more of exploratory and the genes under test have a complicated correlation, then it is more appropriate to use, for example, the Benjamini-Yekutieli(2001) [119] FDR correction.

It has been shown (see section 2.3.4.5) in this thesis that background annotation data greatly affects the enrichment results, and proper pre-processing of the background can lead to a reduction of false positives in the result. This is the case for

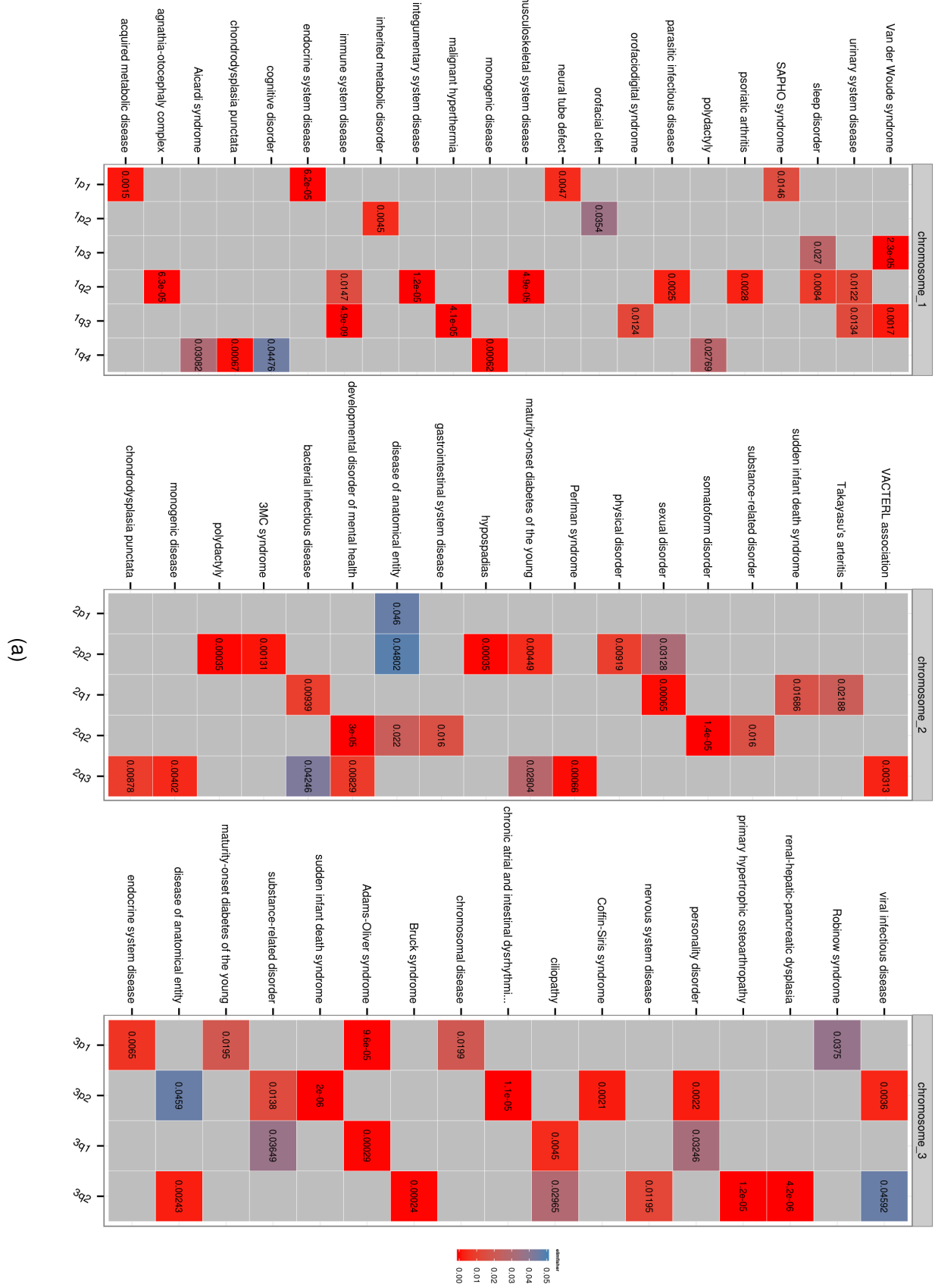
'clip' ontologies where an even more domain specific subset of the ontology is created and used instead of the full ontology. However, the clipping process proposed in [37] requires human experts to manually inspect the ontology and choose appropriate terms, which is difficult and time-consuming. Thus it would be interested to develop automatic methods to achieve the clipping process, which in turns improves the enrichment analysis.

Last but not least, *topOnto* uses topology awareness algorithms to estimate the enrichment p-value, thus it can be computational intensive for certain types of statistic-topology method combinations. For standard over-representation hypergeometric tests, *topOnto* usually returns the result within about a minute depending on the size of the ontology. However, the process can be very slow for the GSEA and GSEA-CSW algorithm when using the *elim* topology method, ranging from hours to days depending on the number of samples and the number of permutations required. This limits the usage of the algorithm despite its advantages at the statistical level, for example, it can not be used on the fly as a web tool. Further optimization is need to speed up the algorithm. One possible approach would be to parallelize part of the processing. The *elim* algorithm considers one level of ontology terms at a time and the result affects the next level of terms, thus within one level, the algorithm is parallelisable. However, the cost of dividing/distributing the task, and merging the results needs to be calculated and tested further to prove that it is possible and efficacious.

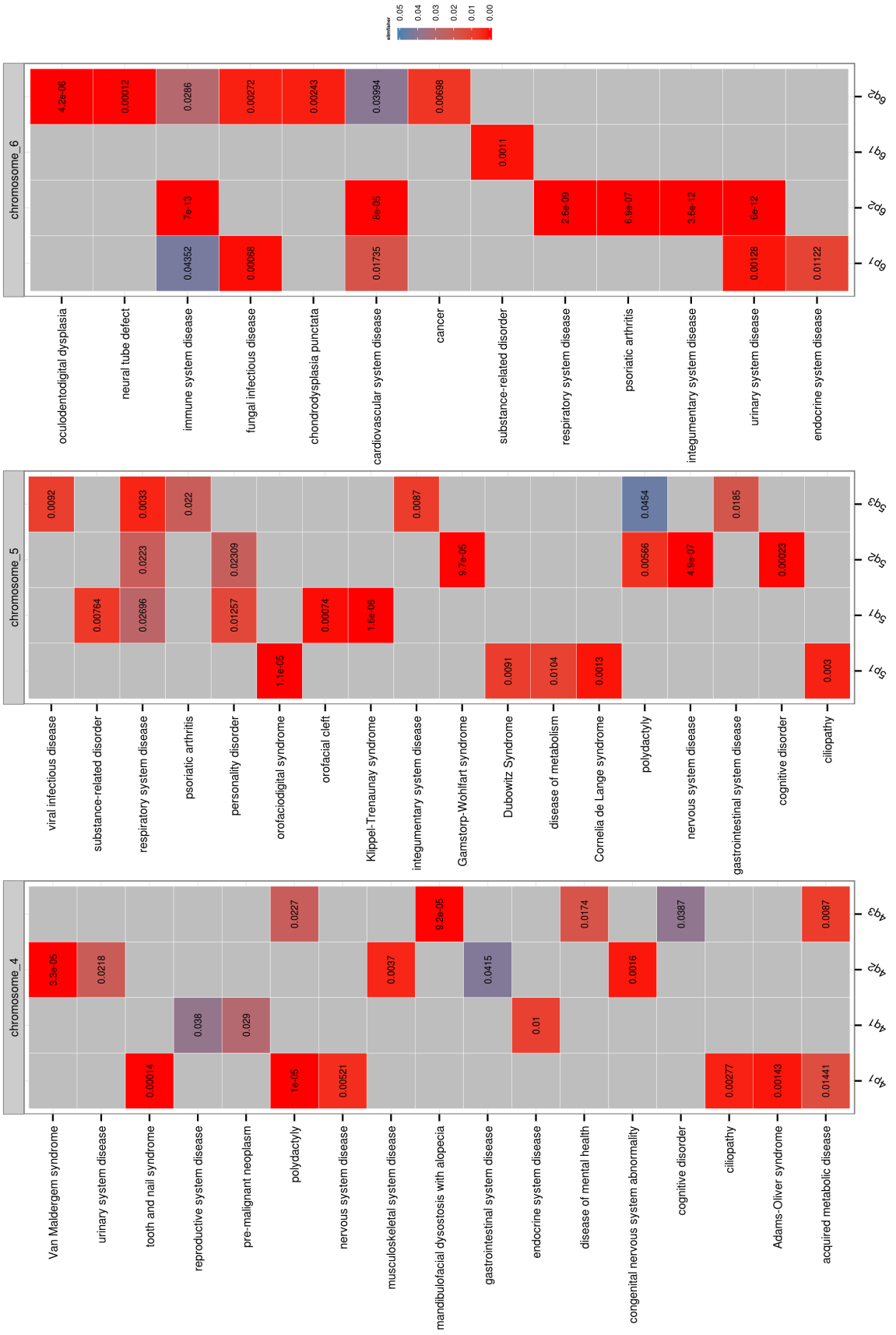
Appendix A

Appendix

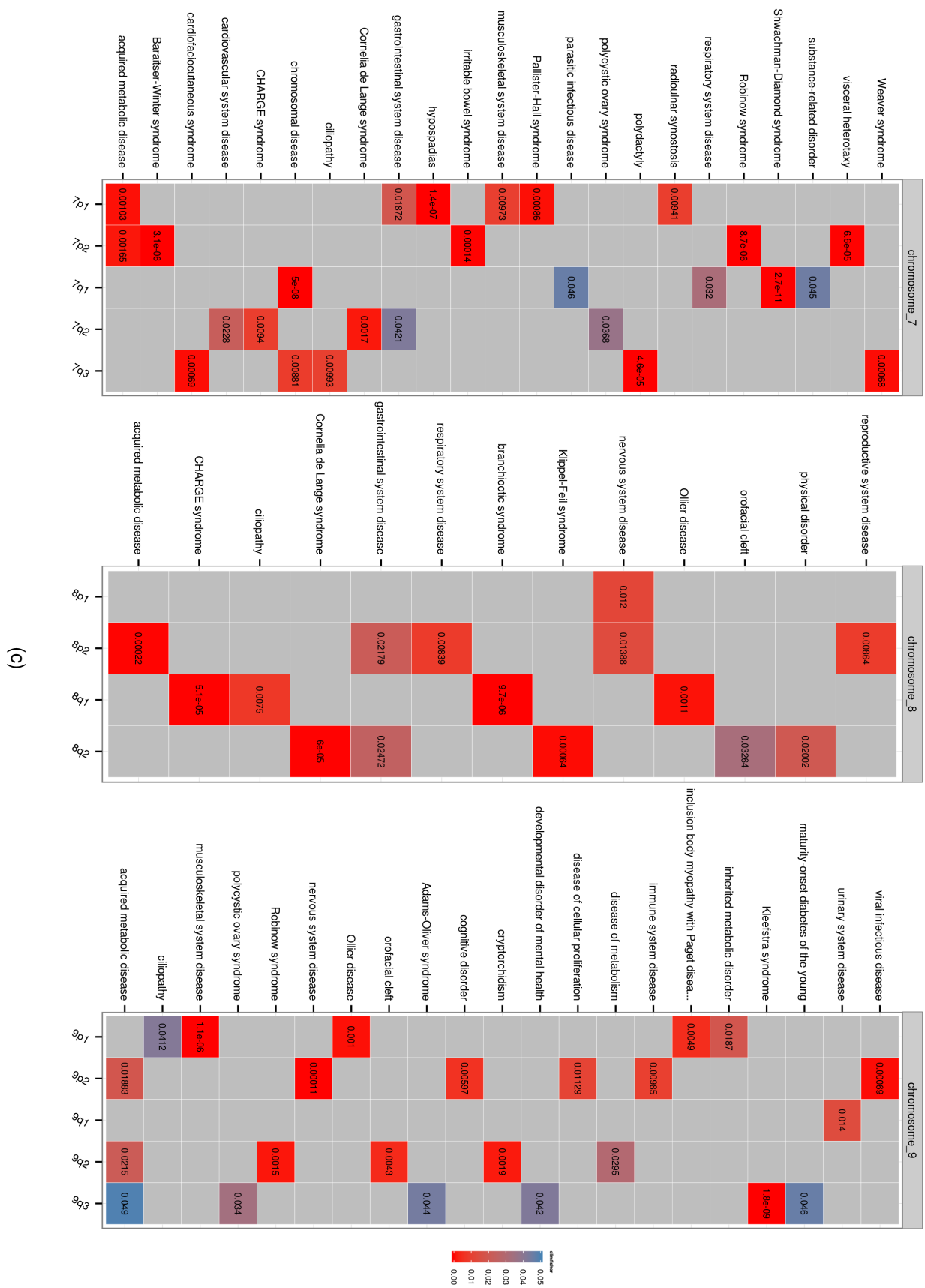
A.1 Chromosomal profiles of disease

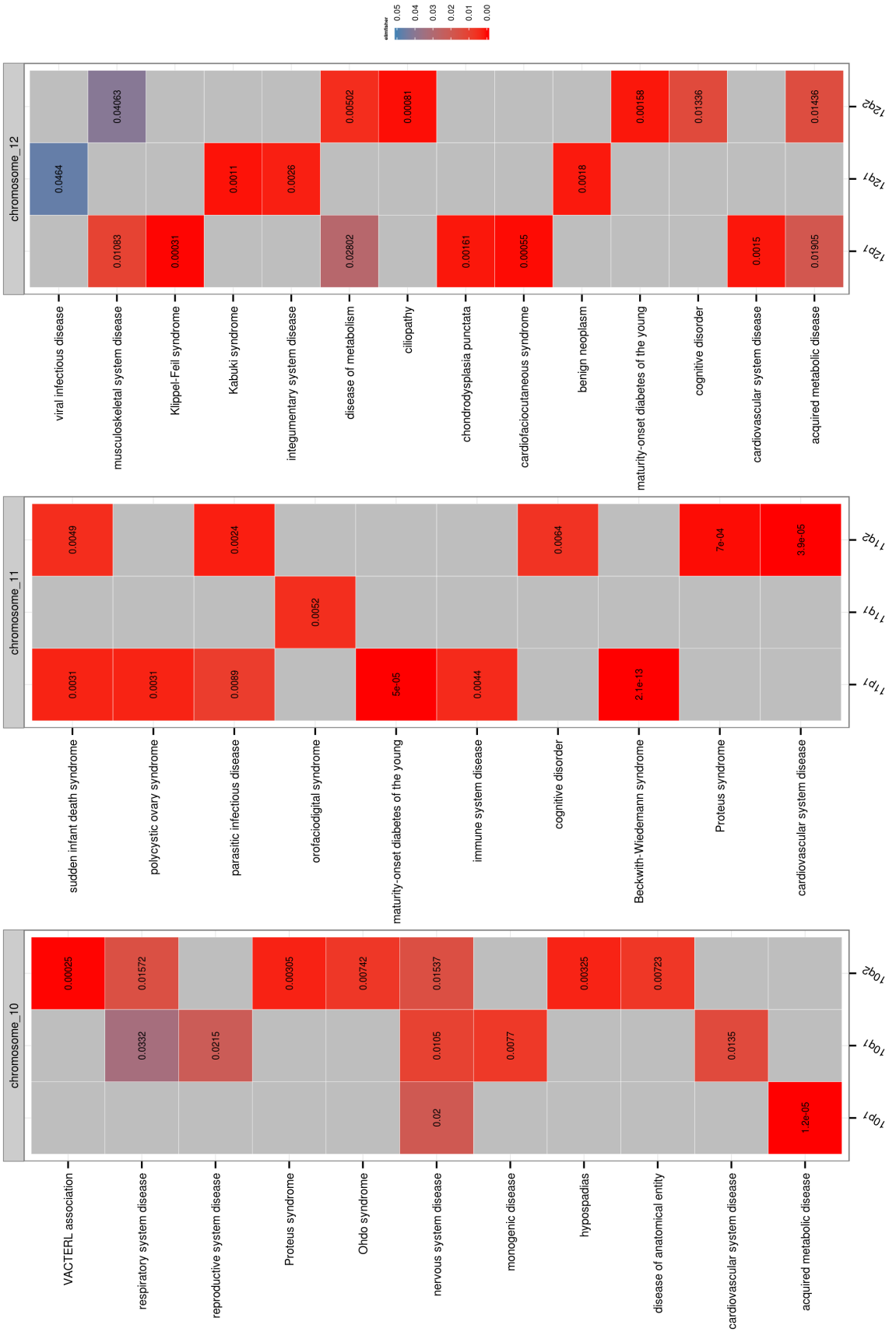

chromosome_1
Van der Woude syndrome
urinary system disease
sleep disorder
SAPHO syndrome
psoriatic arthritis
polydactyly
parasitic infectious disease
orofaciodigital syndrome
orofacial cleft
neural tube defect
musculoskeletal system disease
monogenic disease
malignant hyperthermia
integumentary system disease
inherited metabolic disorder
immune system disease
endocrine system disease
cognitive disorder
chondrodysplasia punctata
Aicardi syndrome
agnathia-otocephaly complex
acquired metabolic disease
1p1
1p2
1p3
1q2
1q3
1q4
chromosome_2
VACTERL association
Takayasu's arteritis
sudden infant death syndrome
substance-related disorder
somatoform disorder
sexual disorder
physical disorder
Perlman syndrome
maturity-onset diabetes of the young
hypospadias
gastrointestinal system disease
disease of anatomical entity
developmental disorder of mental health
bacterial infectious disease
3MC syndrome
polydactyly
monogenic disease
chondrodysplasia punctata
2p1
2p2
2q1
2q2
2q3
chromosome_3
viral infectious disease
Robinow syndrome
renal-hepatic-pancreatic dysplasia
primary hypertrophic osteoarthropathy
personality disorder
nervous system disease
Coffin-Siris syndrome
ciliopathy
chronic atrial and intestinal dysrhythmi...
chromosomal disease
Bruck syndrome
Adams-Oliver syndrome
sudden infant death syndrome
substance-related disorder
maturity-onset diabetes of the young
disease of anatomical entity
endocrine system disease
3p1
3p2
3q1
3q2
0.05
0.04
0.03
0.02
0.01
0.00

(a)

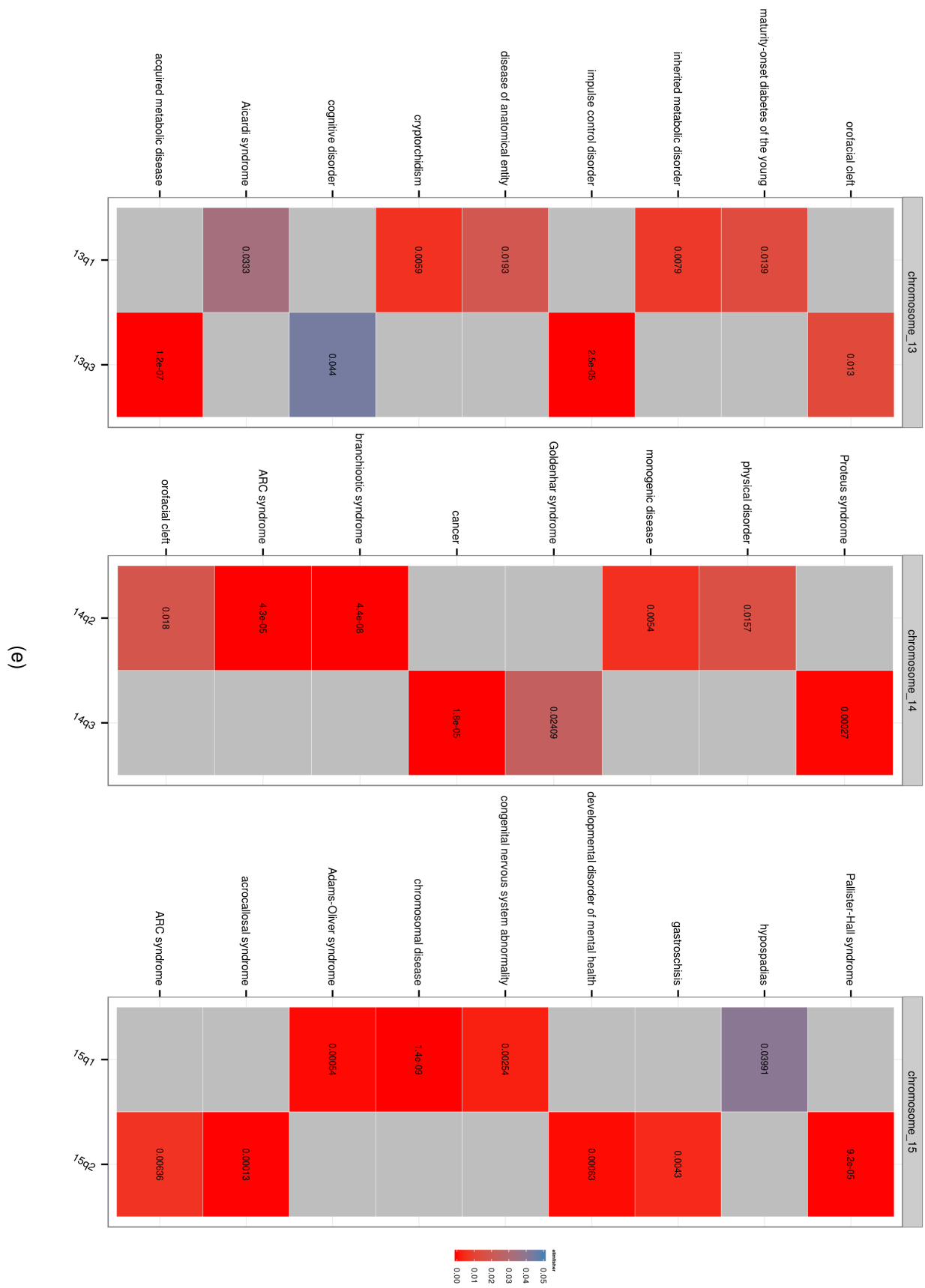

chromosome_4
4p1
4q1
4q2
4q3
Van Maldergem syndrome
urinary system disease
tooth and nail syndrome
reproductive system disease
pre-malignant neoplasm
polydactyly
nervous system disease
musculoskeletal system disease
mandibulofacial dysostosis with alopecia
gastrointestinal system disease
endocrine system disease
disease of mental health
congenital nervous system abnormality
cognitive disorder
ciliopathy
Adams-Oliver syndrome
acquired metabolic disease
3.3e-05
0.0218
0.00014
0.038
0.029
1e-05
0.00521
0.0037
9.2e-05
0.0415
0.01
0.0174
0.0016
0.0387
0.00277
0.00143
0.01441
0.0087
0.0227
chromosome_5
5p1
5q1
5q2
5q3
viral infectious disease
substance-related disorder
respiratory system disease
psoriatic arthritis
personality disorder
orofaciodigital syndrome
orofacial cleft
Klippel-Trenaunay syndrome
integumentary system disease
Gamstorp-Wohlfart syndrome
Dubowitz Syndrome
disease of metabolism
Cornelia de Lange syndrome
polydactyly
nervous system disease
gastrointestinal system disease
cognitive disorder
ciliopathy
0.0092
0.00764
0.02696
0.0223
0.0033
0.022
0.02309
0.01257
1.1e-05
0.00074
1.6e-06
0.0087
9.7e-05
0.0091
0.0104
0.0013
0.0454
0.00566
4.9e-07
0.0185
0.00023
0.003
chromosome_6
6p1
6p2
6q1
6q2
oculodentodigital dysplasia
neural tube defect
immune system disease
fungal infectious disease
chondrodysplasia punctata
cardiovascular system disease
cancer
substance-related disorder
respiratory system disease
psoriatic arthritis
integumentary system disease
urinary system disease
endocrine system disease
4.2e-06
0.00012
0.0286
7e-13
0.00272
0.00243
0.03994
8e-05
0.00698
0.0011
2.6e-09
6.9e-07
3.6e-12
6e-12
0.04352
0.00068
0.01735
0.00128
0.01122
p.adjustvalue
0.05
0.04
0.03
0.02
0.01
0.00

(b)

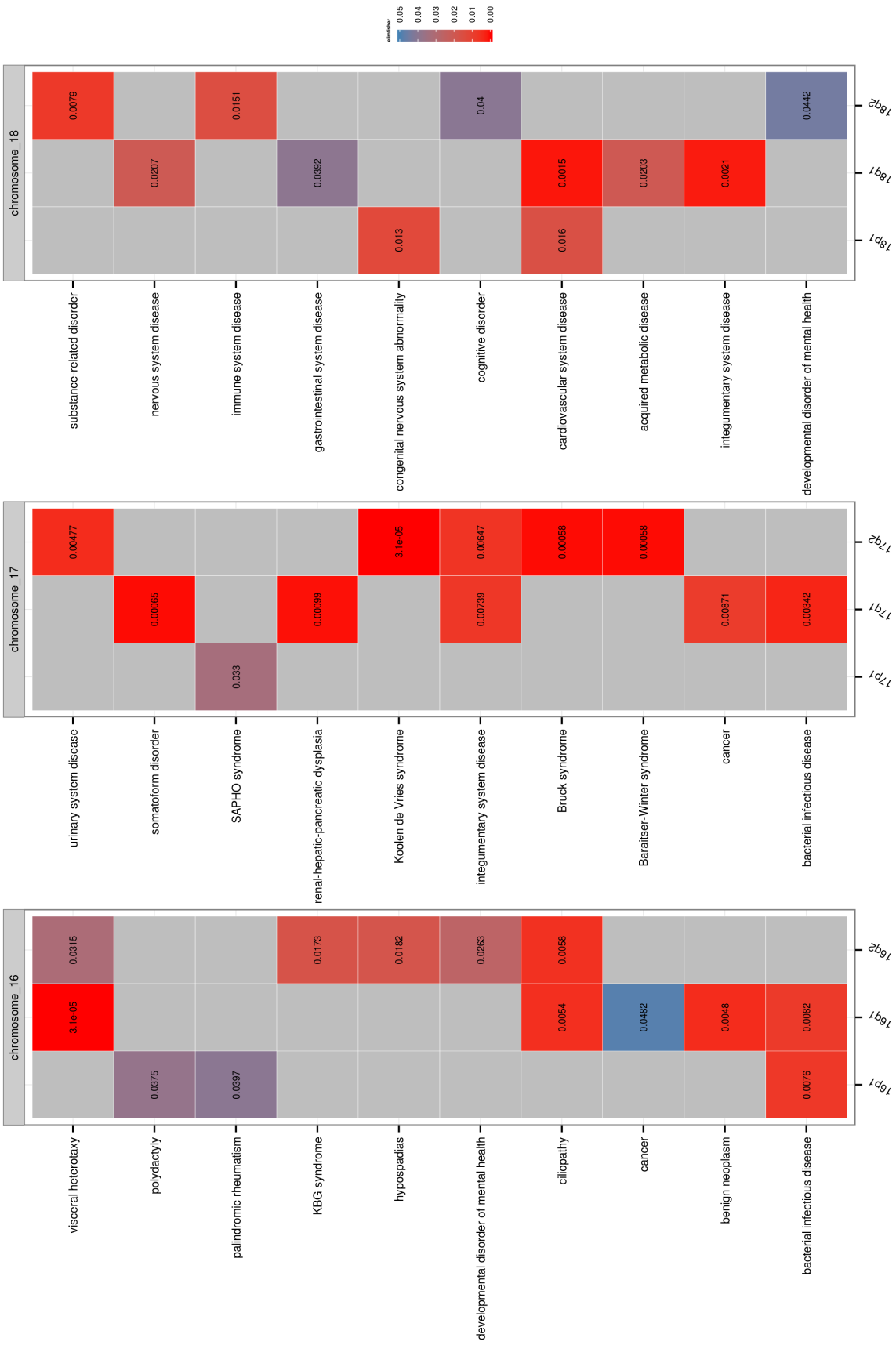

(c)
chromosome_7
Weaver syndrome
visceral heterotaxy
substance-related disorder
Shwachman-Diamond syndrome
Robinow syndrome
respiratory system disease
radioulnar synostosis
polydactyly
polycystic ovary syndrome
parasitic infectious disease
Pallister-Hall syndrome
musculoskeletal system disease
irritable bowel syndrome
hypospadias
gastrointestinal system disease
Cornelia de Lange syndrome
ciliopathy
chromosomal disease
CHARGE syndrome
cardiovascular system disease
cardiofaciocutaneous syndrome
Baraitser-Winter syndrome
acquired metabolic disease
0.00068
6.6e-05
0.045
2.7e-11
8.7e-06
0.032
0.00941
4.6e-05
0.0368
0.046
0.00086
0.00014
0.00973
1.4e-07
0.01872
0.0421
0.0017
0.00993
0.00881
5e-08
0.0094
0.0228
0.00069
3.1e-06
0.00165
0.00103
chromosome_8
reproductive system disease
physical disorder
orofacial cleft
Oliver disease
nervous system disease
Klippel-Feil syndrome
branchiootic syndrome
respiratory system disease
gastrointestinal system disease
Cornelia de Lange syndrome
ciliopathy
CHARGE syndrome
acquired metabolic disease
0.00864
0.02002
0.03264
0.0011
0.012
0.01388
0.00064
9.7e-06
0.00839
0.02179
0.02472
6e-05
0.0075
5.1e-05
0.00022
chromosome_9
viral infectious disease
urinary system disease
maturity-onset diabetes of the young
Kleefstra syndrome
inherited metabolic disorder
inclusion body myopathy with Paget disea...
immune system disease
disease of metabolism
disease of cellular proliferation
developmental disorder of mental health
cryptorchidism
cognitive disorder
Adams-Oliver syndrome
orofacial cleft
Oliver disease
nervous system disease
Robinow syndrome
polycystic ovary syndrome
musculoskeletal system disease
ciliopathy
acquired metabolic disease
0.00069
0.014
0.046
1.8e-09
0.0187
0.0049
0.00985
0.0295
0.01129
0.042
0.0019
0.00597
0.044
0.0043
0.001
0.00011
0.0015
0.034
1.1e-06
0.0412
0.0215
0.049
0.01883
0.05
0.04
0.03
0.02
0.01
0.00


chromosome_10
VACTERL association
respiratory system disease
reproductive system disease
Proteus syndrome
Ohdo syndrome
nervous system disease
monogenic disease
hypospadias
disease of anatomical entity
cardiovascular system disease
acquired metabolic disease
0.00025
0.01572
0.00305
0.00742
0.01537
0.00325
0.00723
0.0332
0.0215
0.0105
0.0077
0.0135
0.02
1.2e-05
10p1
10q1
10q2
chromosome_11
sudden infant death syndrome
polycystic ovary syndrome
parasitic infectious disease
orofaciodigital syndrome
maturity-onset diabetes of the young
immune system disease
cognitive disorder
Beckwith-Wiedemann syndrome
Proteus syndrome
cardiovascular system disease
0.0049
0.0024
0.0064
7e-04
3.9e-05
0.0031
0.0031
0.0089
0.0052
5e-05
0.0044
2.1e-13
11p1
11q1
11q2
chromosome_12
viral infectious disease
musculoskeletal system disease
Klippel-Feil syndrome
Kabuki syndrome
integumentary system disease
disease of metabolism
ciliopathy
chondrodysplasia punctata
cardiofaciocutaneous syndrome
benign neoplasm
maturity-onset diabetes of the young
cognitive disorder
cardiovascular system disease
acquired metabolic disease
0.04063
0.00502
0.00081
0.00158
0.01336
0.01436
0.0464
0.0011
0.0026
0.0018
0.01083
0.00031
0.02802
0.00161
0.00055
0.0015
0.01905
12p1
12q1
12q2
0.05
0.04
0.03
0.02
0.01
0.00

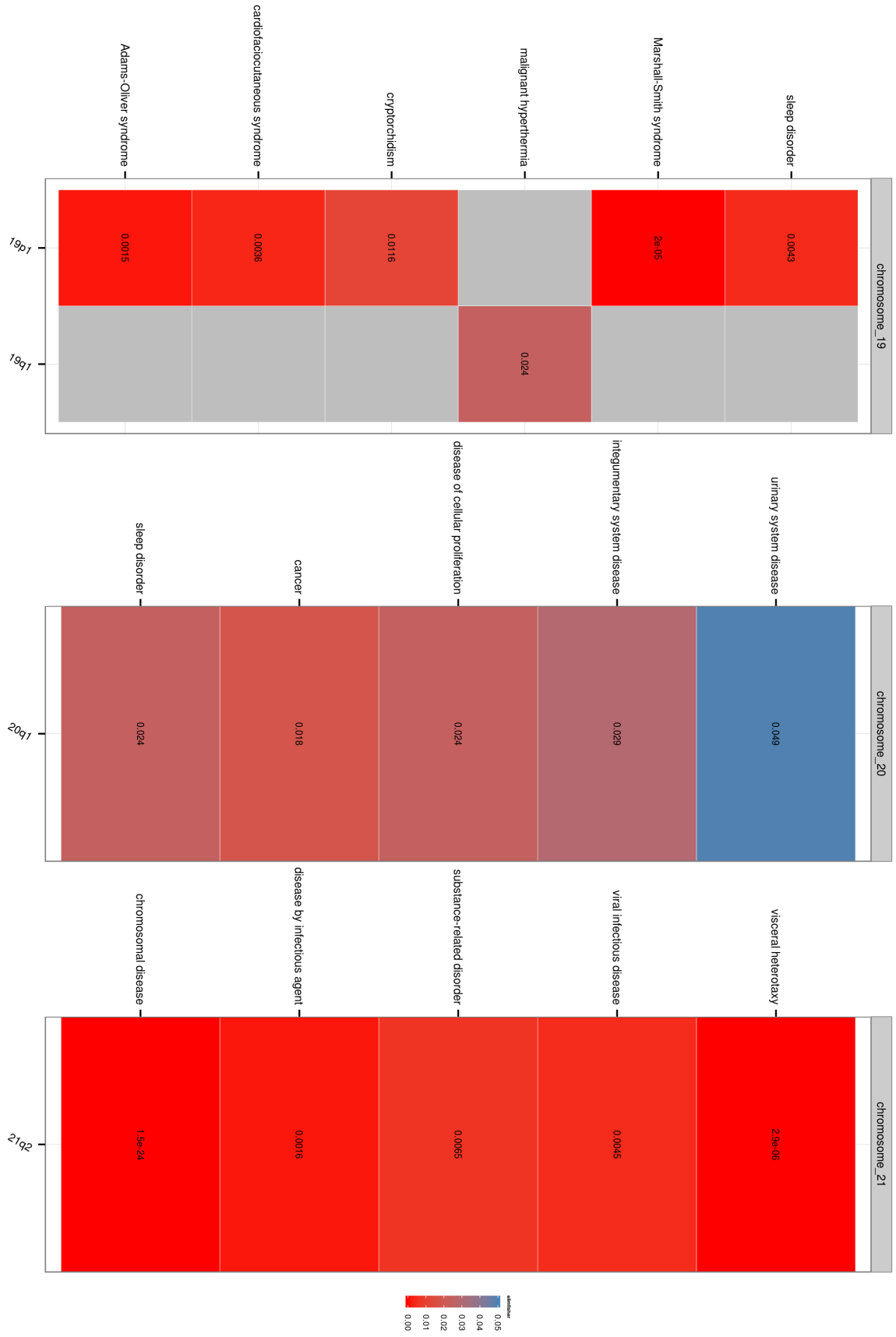
(d)


chromosome_13
orofacial cleft
maturity-onset diabetes of the young
inherited metabolic disorder
impulse control disorder
disease of anatomical entity
cryptorchidism
cognitive disorder
Alcardi syndrome
acquired metabolic disease
13q1
13q3
0.013
0.0139
0.0079
2.5e-05
0.0193
0.0059
0.044
0.0333
1.2e-07
chromosome_14
Proteus syndrome
physical disorder
monogenic disease
Goldenhar syndrome
cancer
branchiootic syndrome
ARC syndrome
orofacial cleft
14q2
14q3
0.00027
0.0157
0.0054
0.02409
1.8e-05
4.4e-08
4.3e-05
0.018
chromosome_15
Pallister-Hall syndrome
hypospadias
gastroschisis
developmental disorder of mental health
congenital nervous system abnormality
chromosomal disease
Adams-Oliver syndrome
acrocallosal syndrome
ARC syndrome
15q1
15q2
9.2e-05
0.03991
0.0043
0.00063
0.00254
1.4e-09
0.00054
0.00013
0.00636
adjusted p
0.05
0.04
0.03
0.02
0.01
0.00

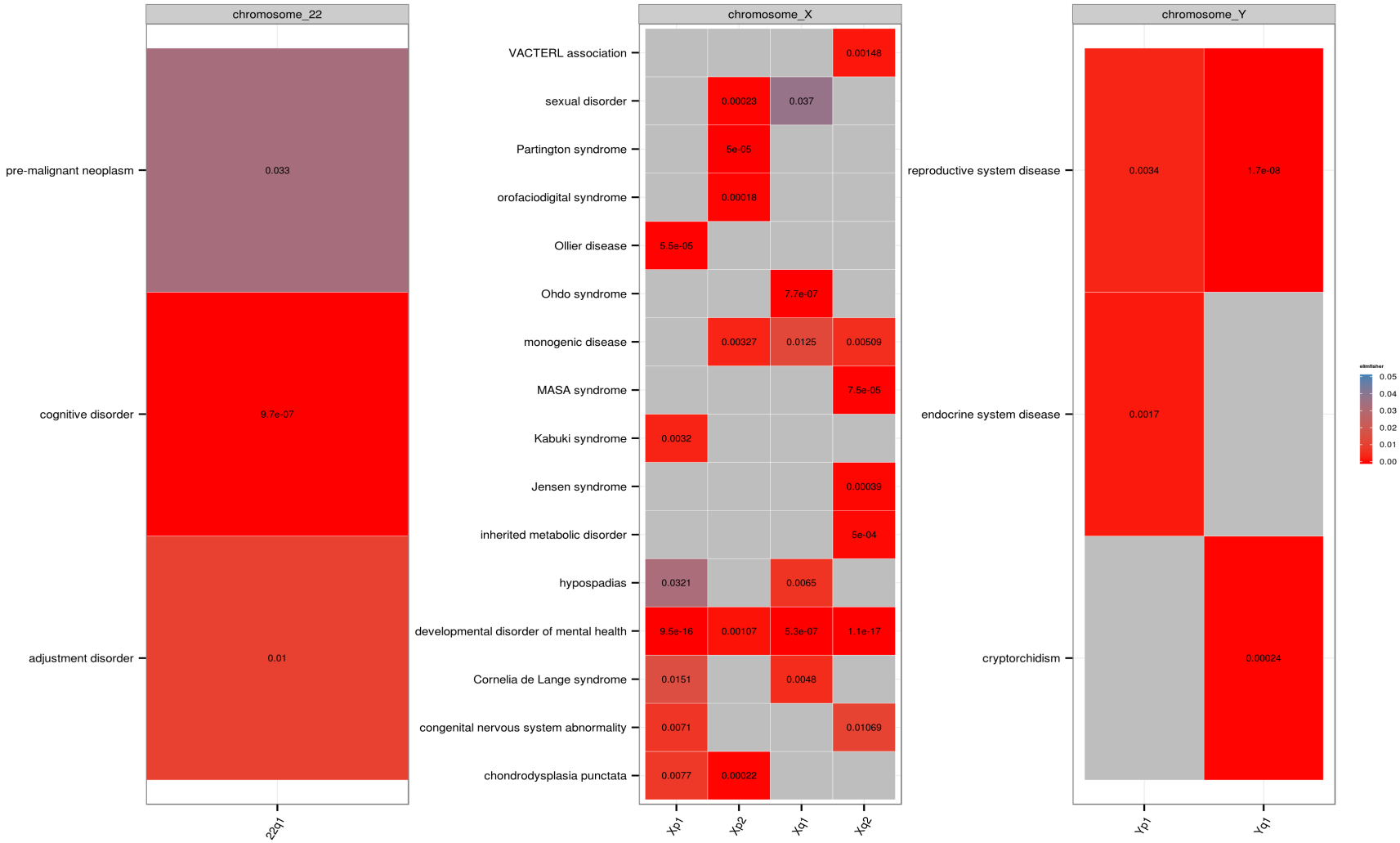
(e)


chromosome_16
visceral heterotaxy
polydactyly
palindromic rheumatism
KBG syndrome
hypospadias
developmental disorder of mental health
ciliopathy
cancer
benign neoplasm
bacterial infectious disease
0.0315
3.1e-05
0.0375
0.0397
0.0173
0.0182
0.0263
0.0058
0.0054
0.0482
0.0048
0.0082
0.0076
16p1
16q1
16q2
chromosome_17
urinary system disease
somatoform disorder
SAPHO syndrome
renal-hepatic-pancreatic dysplasia
Koolen de Vries syndrome
integumentary system disease
Bruck syndrome
Baraitser-Winter syndrome
cancer
bacterial infectious disease
0.00477
0.00065
0.033
0.00099
3.1e-05
0.00647
0.00739
0.00058
0.00058
0.00871
0.00342
17p1
17q1
17q2
chromosome_18
substance-related disorder
nervous system disease
immune system disease
gastrointestinal system disease
congenital nervous system abnormality
cognitive disorder
cardiovascular system disease
acquired metabolic disease
integumentary system disease
developmental disorder of mental health
0.0079
0.0207
0.0151
0.0392
0.013
0.04
0.0015
0.016
0.0203
0.0021
0.0442
18p1
18q1
18q2
0.05
0.04
0.03
0.02
0.01
0.00

(f)


chromosome_19
sleep disorder
Marshall-Smith syndrome
malignant hyperthermia
cryptorchidism
cardiofaciocutaneous syndrome
Adams-Oliver syndrome
0.0043
2e-05
0.024
0.0116
0.0036
0.0015
19p1
19q1
chromosome_20
urinary system disease
integumentary system disease
disease of cellular proliferation
cancer
sleep disorder
0.049
0.029
0.024
0.018
0.024
20q1
chromosome_21
visceral heterotaxy
viral infectious disease
substance-related disorder
disease by infectious agent
chromosomal disease
2.9e-06
0.0045
0.0065
0.0016
1.5e-24
21q2
0.05
0.04
0.03
0.02
0.01
0.00

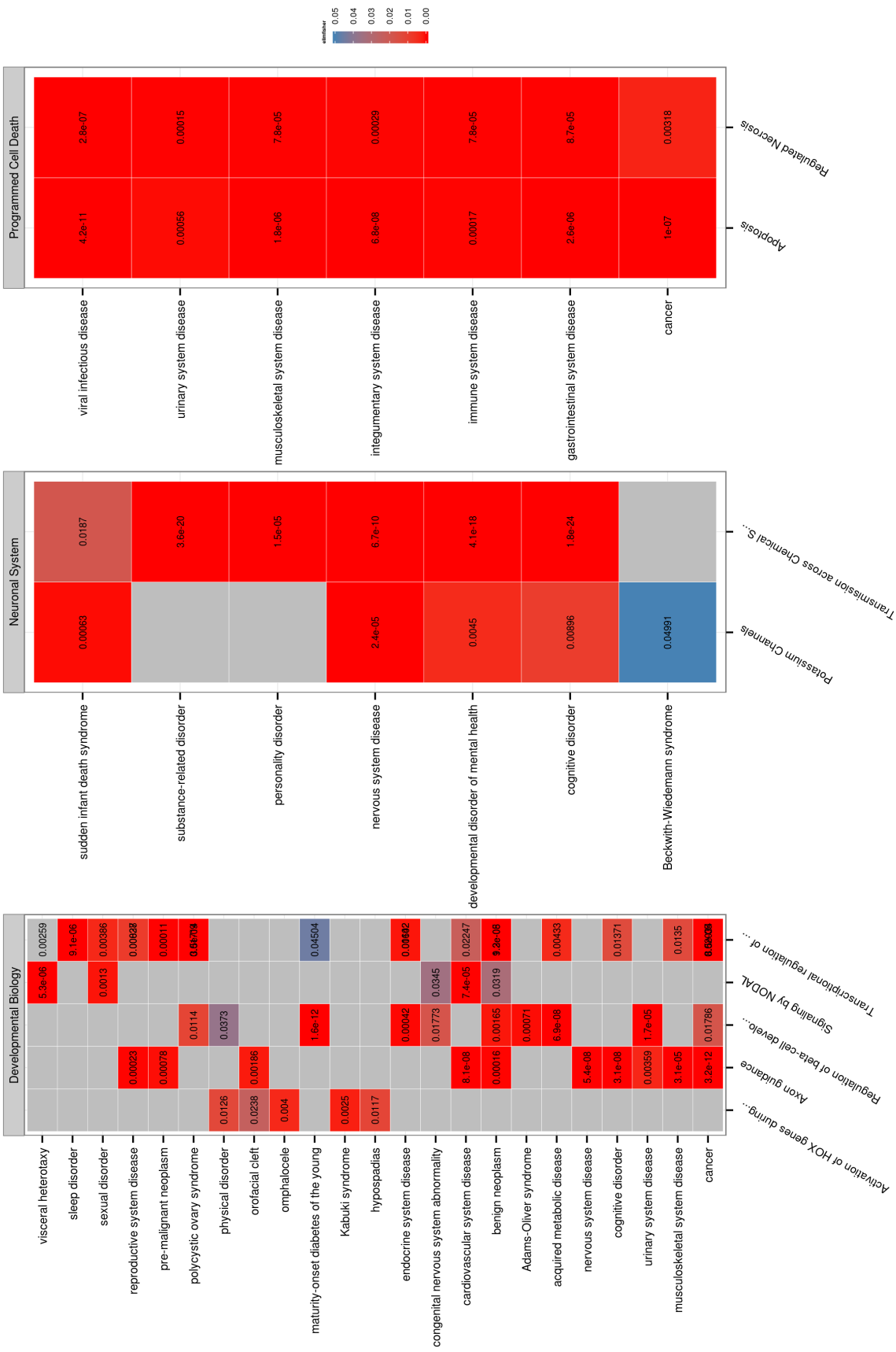
(g)



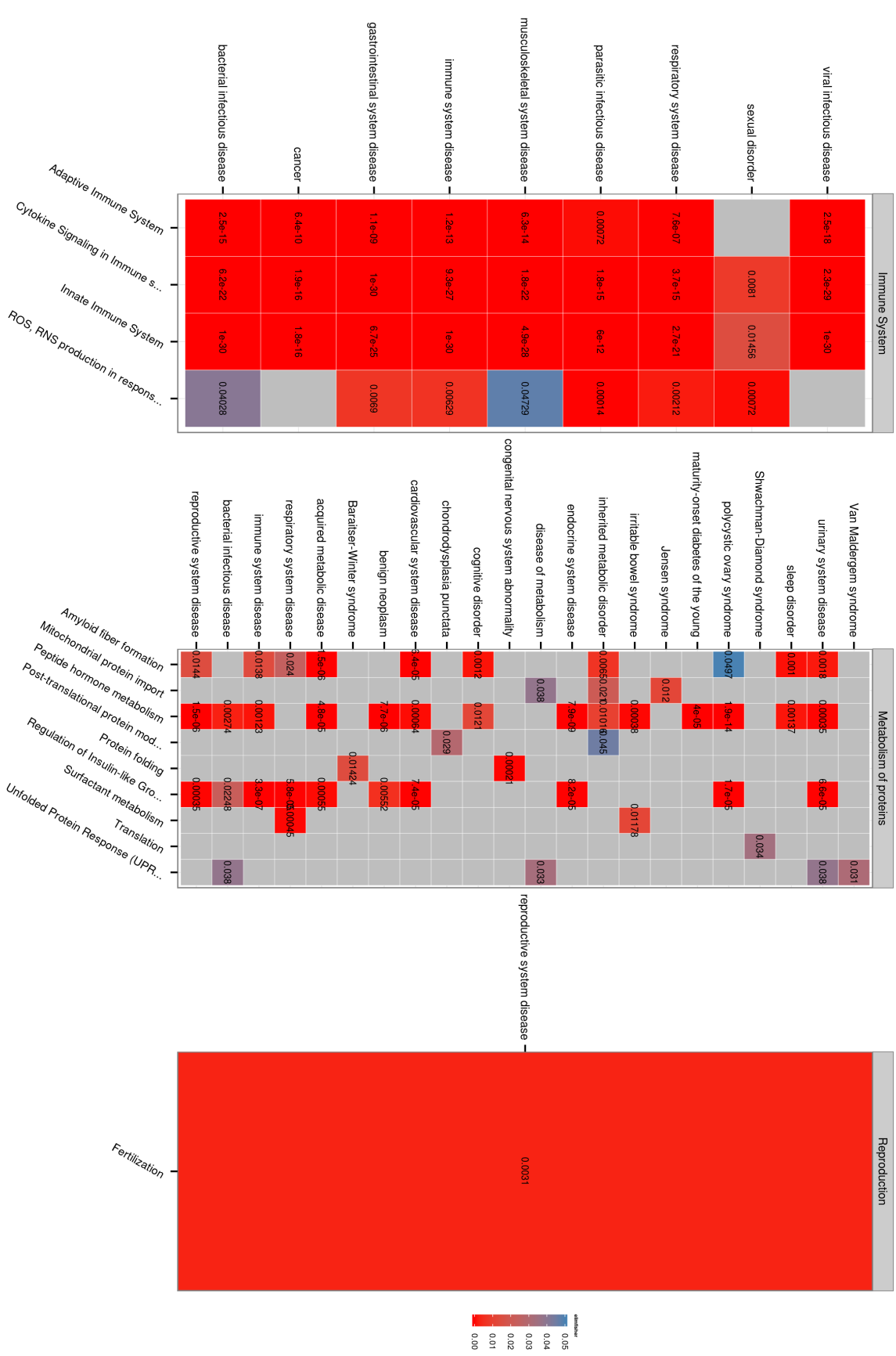
(h)

Figure A1: Heat map of disease profile of 76 level 2 Chromosome Location(one sub band) grouped by 24 top level CO terms. Disease enrichment analysis were perform using *topOnto* and *hdgdb* with the *elim* topology methods. 136 level 3 HDO terms were selected to represent human disease and the top 5 enriched diseases for each CO term were shown in the Figure with the corresponding p-value.

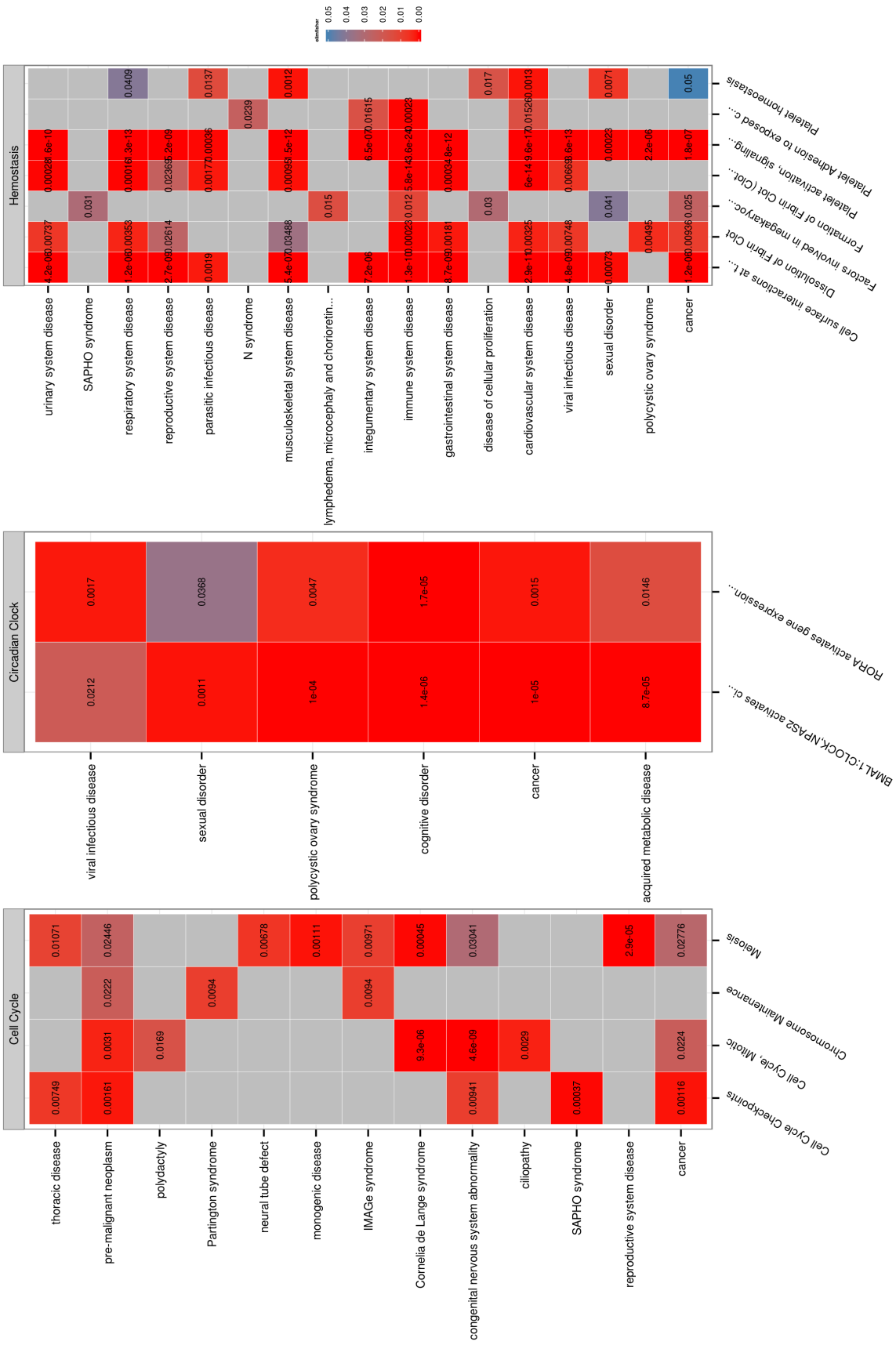
A.2 Reactome pathway profiles of disease

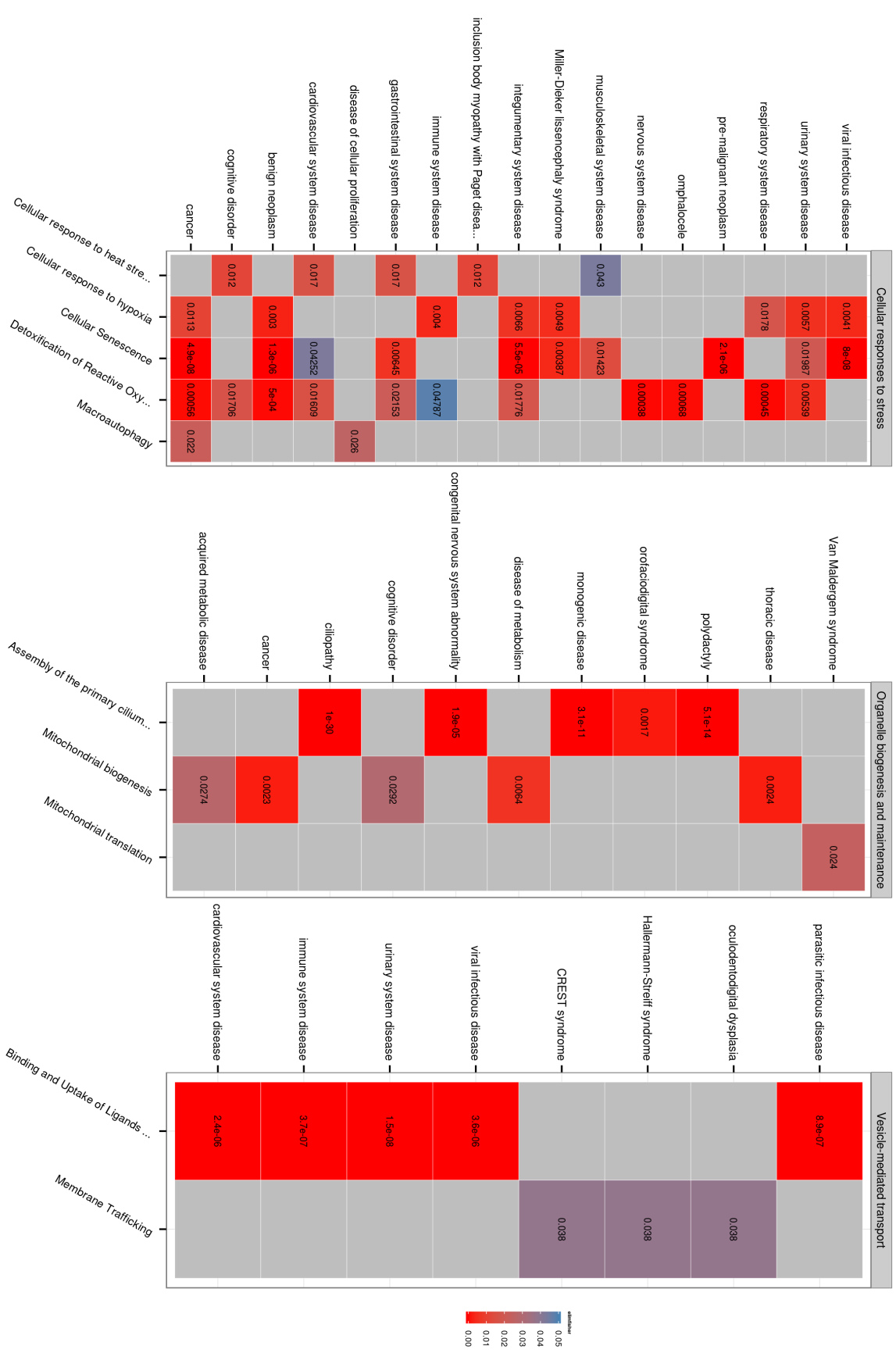

Developmental Biology
Neuronal System
Programmed Cell Death
visceral heterotaxy
sleep disorder
sexual disorder
reproductive system disease
pre-malignant neoplasm
polycystic ovary syndrome
physical disorder
orofacial cleft
omphalocele
maturity-onset diabetes of the young
Kabuki syndrome
hypospadias
endocrine system disease
congenital nervous system abnormality
cardiovascular system disease
benign neoplasm
Adams-Oliver syndrome
acquired metabolic disease
nervous system disease
cognitive disorder
urinary system disease
musculoskeletal system disease
cancer
Activation of HOX genes during...
Axon guidance
Regulation of beta-cell develo...
Signaling by NODAL
Transcriptional regulation of ...
sudden infant death syndrome
substance-related disorder
personality disorder
nervous system disease
developmental disorder of mental health
cognitive disorder
Beckwith-Wiedemann syndrome
Potassium Channels
Transmission across Chemical S...
viral infectious disease
urinary system disease
musculoskeletal system disease
integumentary system disease
immune system disease
gastrointestinal system disease
cancer
Apoptosis
Regulated Necrosis
adjusted p.value
0.05
0.04
0.03
0.02
0.01
0.00

(a)

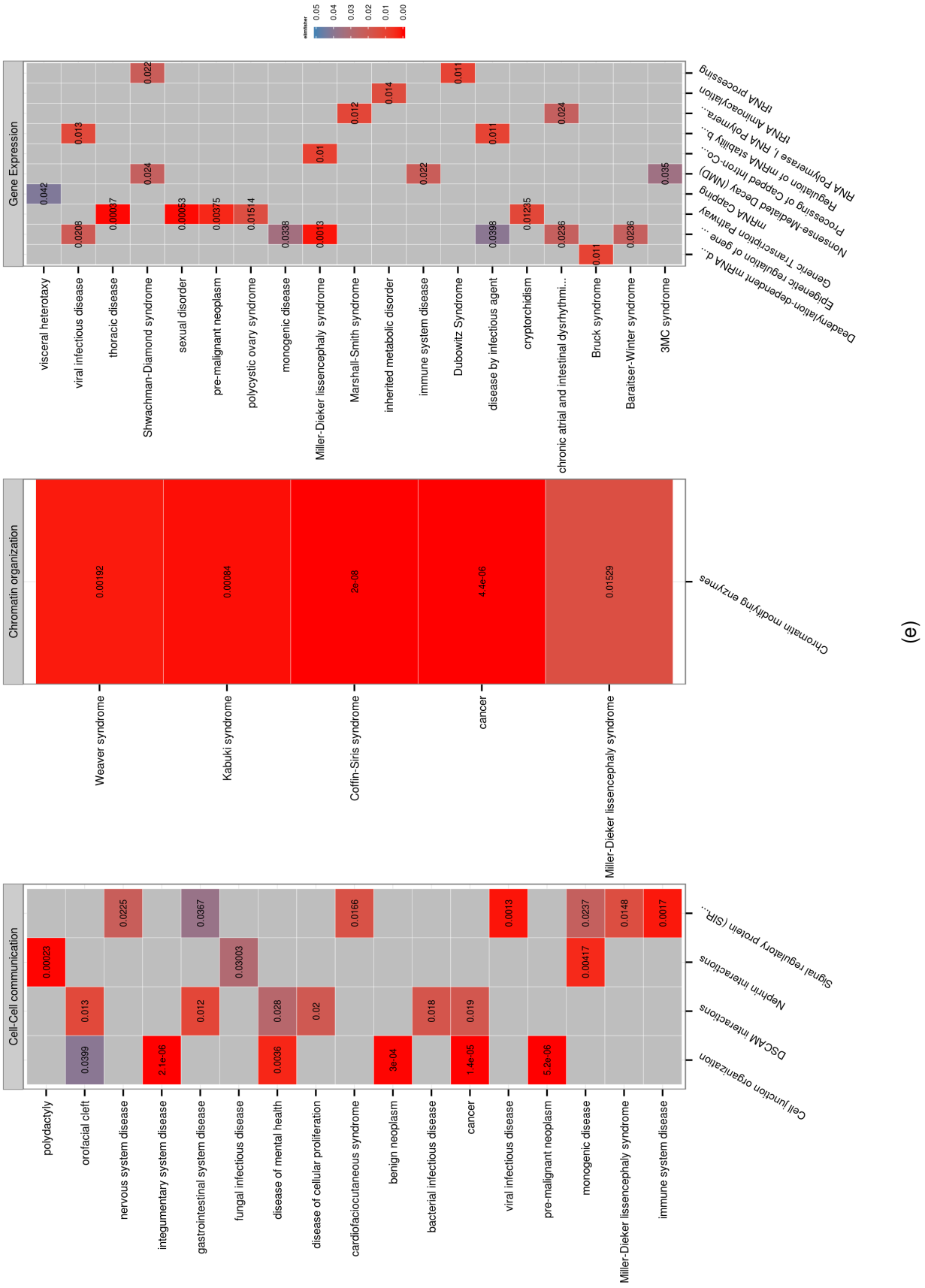

Immune System
Metabolism of proteins
Reproduction
viral infectious disease
sexual disorder
respiratory system disease
parasitic infectious disease
musculoskeletal system disease
immune system disease
gastrointestinal system disease
cancer
bacterial infectious disease
Van Maldergem syndrome
urinary system disease
sleep disorder
Shwachman-Diamond syndrome
polycystic ovary syndrome
maturity-onset diabetes of the young
Jensen syndrome
irritable bowel syndrome
inherited metabolic disorder
endocrine system disease
disease of metabolism
congenital nervous system abnormality
cognitive disorder
chondrodysplasia punctata
cardiovascular system disease
benign neoplasm
Baraitser-Winter syndrome
acquired metabolic disease
respiratory system disease
immune system disease
bacterial infectious disease
reproductive system disease
reproductive system disease
Fertilization
Adaptive Immune System
Cytokine Signaling in Immune s...
Innate Immune System
ROS, RNS production in respons...
Amyloid fiber formation
Mitochondrial protein import
Peptide hormone metabolism
Post-translational protein mod...
Protein folding
Regulation of Insulin-like Gro...
Surfactant metabolism
Translation
Unfolded Protein Response (UPR...
estimate
0.05
0.04
0.03
0.02
0.01
0.00

(b)

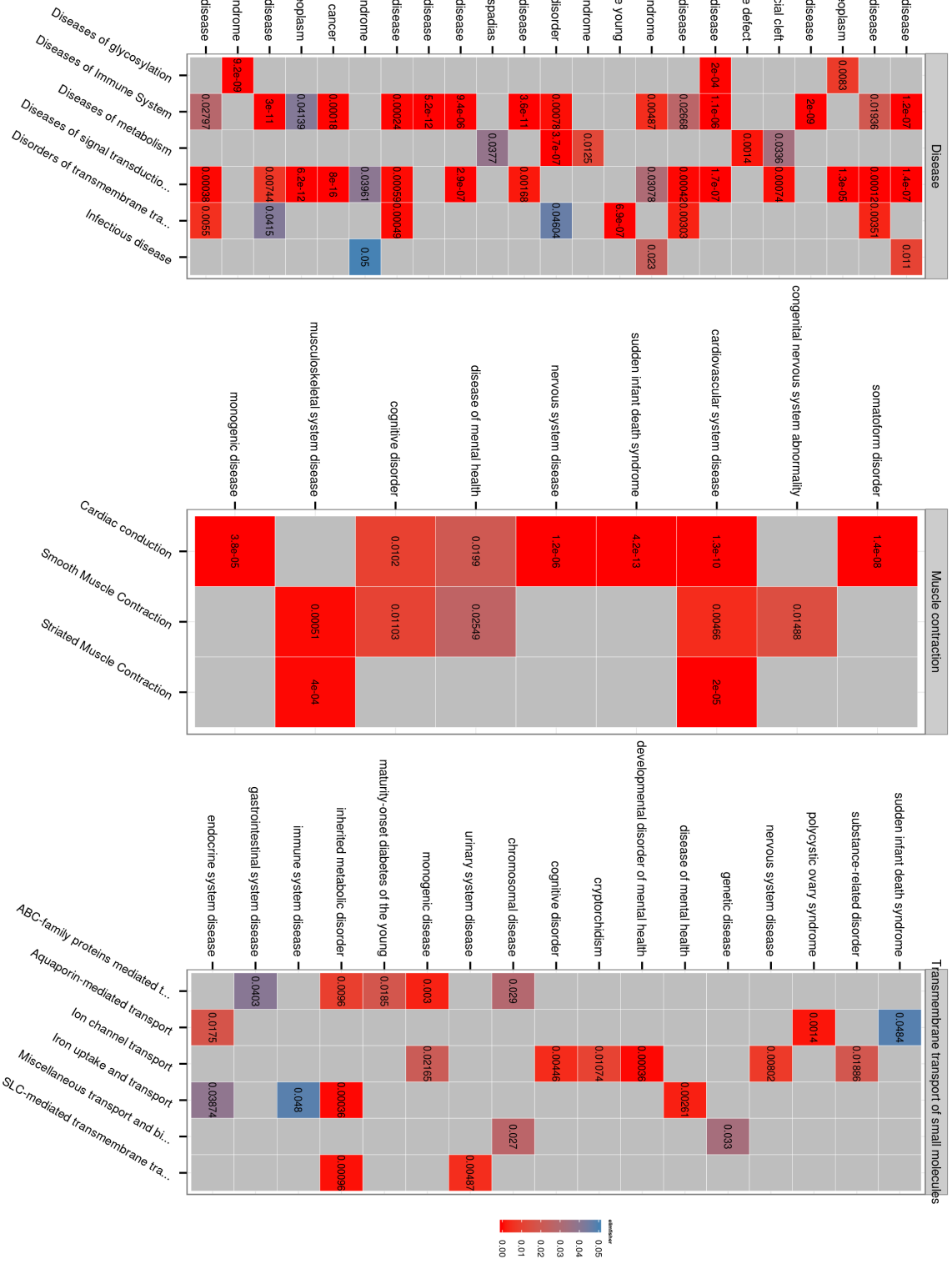

Cell Cycle
Circadian Clock
Hemostasis
thoracic disease
pre-malignant neoplasm
polydactyly
Partington syndrome
neural tube defect
monogenic disease
IMAGe syndrome
Cornelia de Lange syndrome
congenital nervous system abnormality
ciliopathy
SAPHO syndrome
reproductive system disease
cancer
Cell Cycle Checkpoints
Cell Cycle, Mitotic
Chromosome Maintenance
Meiosis
viral infectious disease
sexual disorder
polycystic ovary syndrome
cognitive disorder
cancer
acquired metabolic disease
BMAL1:CLOCK,NPAS2 activates cl...
RORA activates gene expression...
urinary system disease
SAPHO syndrome
respiratory system disease
reproductive system disease
parasitic infectious disease
N syndrome
musculoskeletal system disease
lymphedema, microcephaly and chorioretin...
integumentary system disease
immune system disease
gastrointestinal system disease
disease of cellular proliferation
cardiovascular system disease
viral infectious disease
sexual disorder
polycystic ovary syndrome
cancer
Cell surface interactions at t...
Dissolution of Fibrin Clot
Factors involved in megakaryoc...
Formation of Fibrin Clot (Clot...
Platelet activation, signaling...
Platelet Adhesion to exposed c...
Platelet homeostasis
(c)


Cellular responses to stress
viral infectious disease
urinary system disease
respiratory system disease
pre-malignant neoplasm
omphalocele
nervous system disease
musculoskeletal system disease
Miller-Dieker lissencephaly syndrome
integumentary system disease
inclusion body myopathy with Paget disea...
immune system disease
gastrointestinal system disease
disease of cellular proliferation
cardiovascular system disease
benign neoplasm
cognitive disorder
cancer
Cellular response to heat stre...
Cellular response to hypoxia
Cellular Senescence
Detoxification of Reactive Oxy...
Macroautophagy
0.0041 8e-08
0.0057 0.01987 0.00539
0.0178 0.00045
2.1e-06
0.00068
0.00038
0.043 0.01423
0.0049 0.00387
0.0066 5.5e-05 0.01776
0.004 0.04787
0.012
0.017 0.00645 0.02153
0.026
0.017 0.04252 0.01609
0.003 1.3e-06 5e-04
0.012 0.01706
0.0113 4.9e-08 0.00056 0.022
Organelle biogenesis and maintenance
Van Maldergem syndrome
thoracic disease
polydactyly
oculodentodigital dysplasia
Hallermann-Streiff syndrome
CREST syndrome
viral infectious disease
urinary system disease
immune system disease
cardiovascular system disease
parasitic infectious disease
0.024
0.0024
5.1e-14
0.0017
3.1e-11
0.0064
1.9e-05
0.0292
1e-30
0.0023
0.0274
Vesicle-mediated transport
congenital nervous system abnormality
disease of metabolism
monogenic disease
orofaciodigital syndrome
cognitive disorder
ciliopathy
cancer
acquired metabolic disease
Assembly of the primary cilium...
Mitochondrial biogenesis
Mitochondrial translation
8.9e-07
0.038
3.6e-06
0.038
1.5e-08
0.038
3.7e-07
0.038
2.4e-06
Binding and Uptake of Ligands ...
Membrane Trafficking
0.00
0.01
0.02
0.03
0.04
0.05

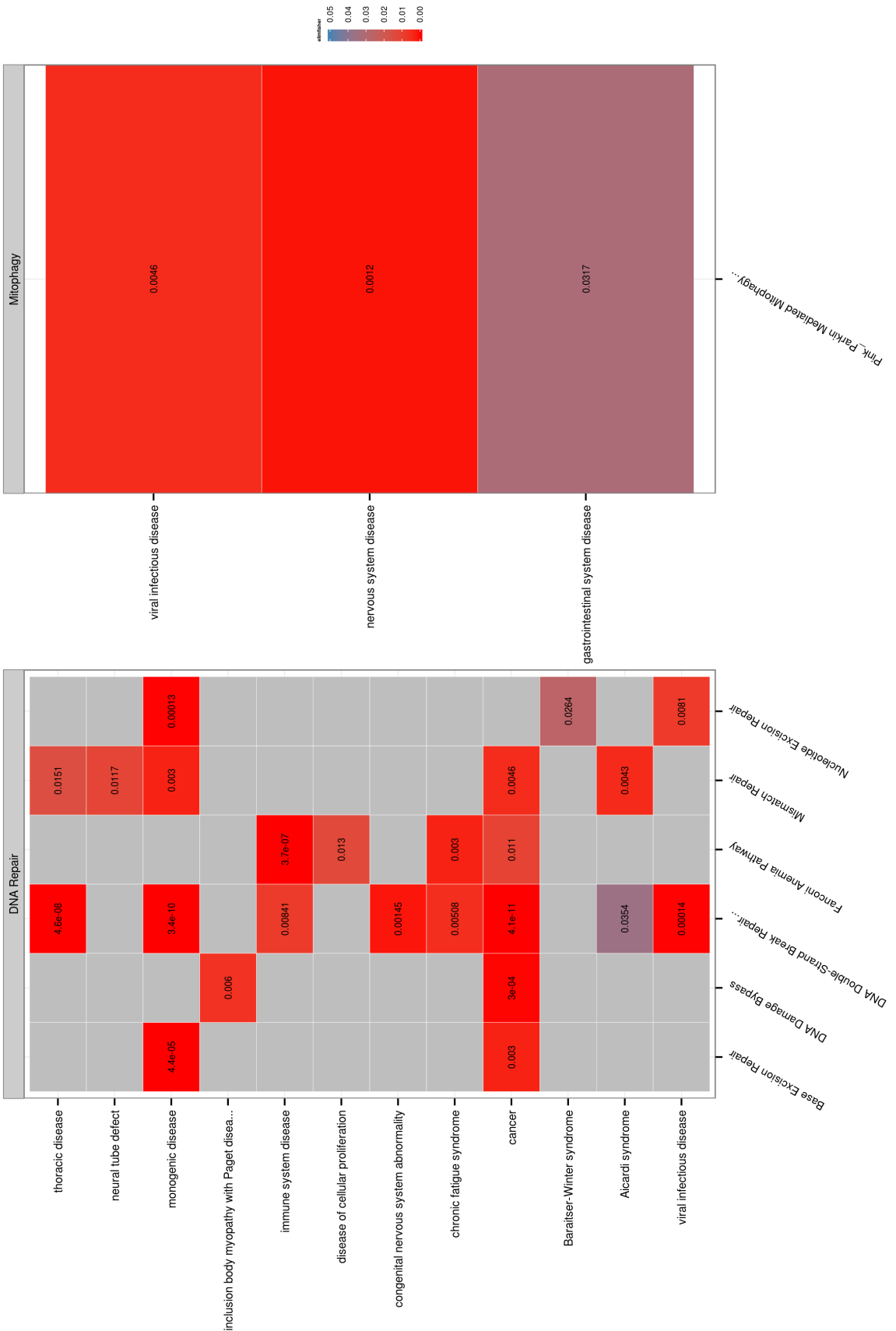
(d)


Gene Expression
Chromatin organization
Cell-Cell communication
visceral heterotaxy
viral infectious disease
thoracic disease
Shwachman-Diamond syndrome
sexual disorder
pre-malignant neoplasm
polycystic ovary syndrome
monogenic disease
Miller-Dieker lissencephaly syndrome
Marshall-Smith syndrome
inherited metabolic disorder
immune system disease
Dubowitz Syndrome
disease by infectious agent
cryptorchidism
chronic atrial and intestinal dysrhythmi...
Bruck syndrome
Baraitser-Winter syndrome
3MC syndrome
Deadenylation-dependent mRNA d...
Epigenetic regulation of gene ...
Generic Transcription Pathway
mRNA Capping
Nonsense-Mediated Decay (NMD)
Processing of Capped Intron-Co...
Regulation of mRNA stability b...
RNA Polymerase I, RNA Polymera...
tRNA Aminoacylation
tRNA processing
Weaver syndrome
Kabuki syndrome
Coffin-Siris syndrome
cancer
Miller-Dieker lissencephaly syndrome
Chromatin modifying enzymes
polydactyly
orofacial cleft
nervous system disease
integumentary system disease
gastrointestinal system disease
fungal infectious disease
disease of mental health
disease of cellular proliferation
cardiofaciocutaneous syndrome
benign neoplasm
bacterial infectious disease
cancer
viral infectious disease
pre-malignant neoplasm
monogenic disease
Miller-Dieker lissencephaly syndrome
immune system disease
Cell junction organization
DSCAM interactions
Nephrin interactions
Signal regulatory protein (SIR...

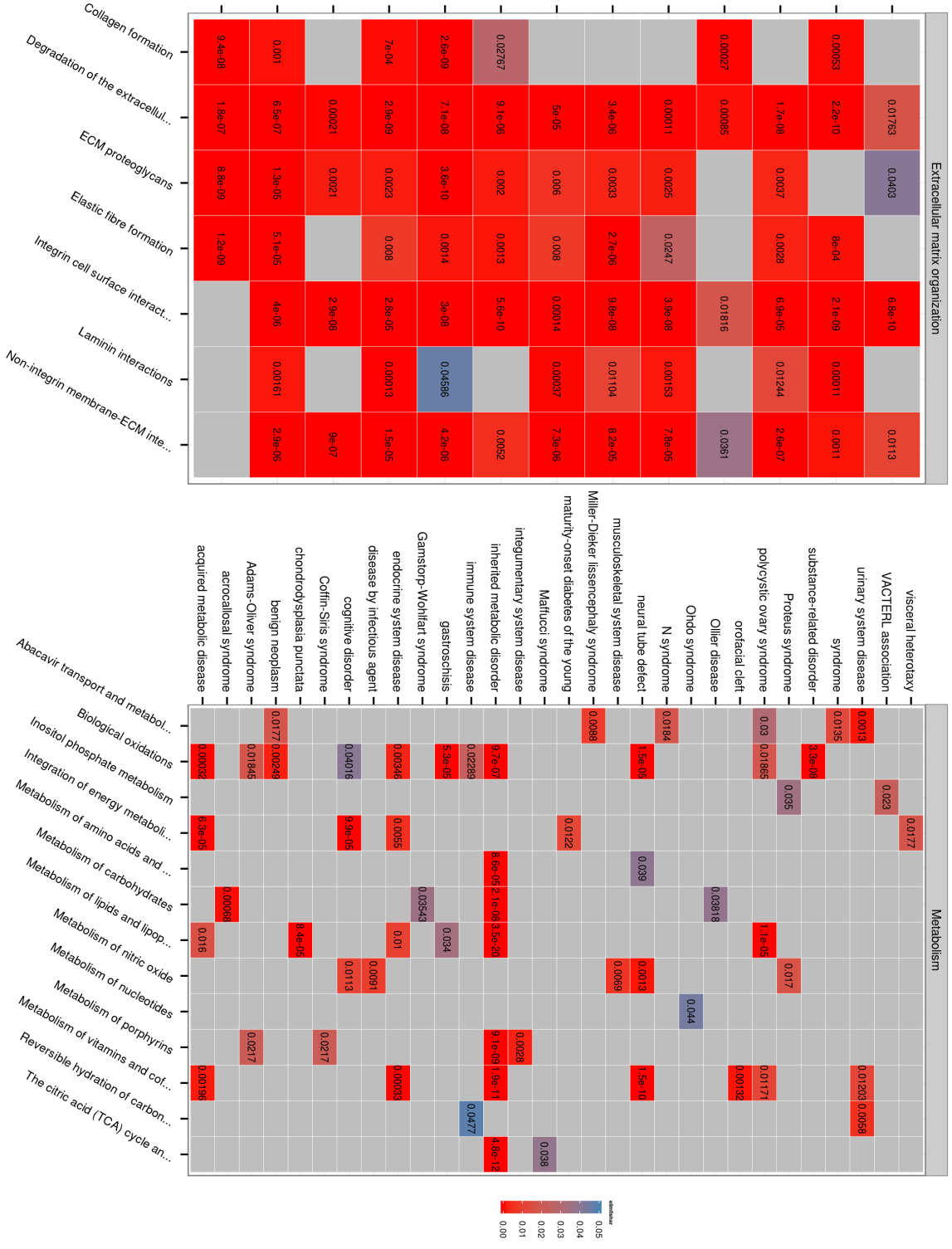
(e)


Disease
viral infectious disease
urinary system disease
pre-malignant neoplasm
parasitic infectious disease
orofacial cleft
neural tube defect
musculoskeletal system disease
monogenic disease
Miller-Dieker lissencephaly syndrome
maturity-onset diabetes of the young
Maffucci syndrome
inherited metabolic disorder
immune system disease
hypospadias
gastrointestinal system disease
fungal infectious disease
endocrine system disease
cardiofaciocutaneous syndrome
cancer
benign neoplasm
bacterial infectious disease
acrocallosal syndrome
acquired metabolic disease
Diseases of glycosylation
Diseases of Immune System
Diseases of metabolism
Diseases of signal transductio...
Disorders of transmembrane tra...
Infectious disease
Muscle contraction
somatoform disorder
congenital nervous system abnormality
cardiovascular system disease
sudden infant death syndrome
nervous system disease
disease of mental health
cognitive disorder
musculoskeletal system disease
monogenic disease
Cardiac conduction
Smooth Muscle Contraction
Striated Muscle Contraction
Transmembrane transport of small molecules
sudden infant death syndrome
substance-related disorder
polycystic ovary syndrome
nervous system disease
genetic disease
disease of mental health
developmental disorder of mental health
cryptorchidism
cognitive disorder
chromosomal disease
urinary system disease
monogenic disease
maturity-onset diabetes of the young
inherited metabolic disorder
immune system disease
gastrointestinal system disease
endocrine system disease
ABC-family proteins mediated t...
Aquaporin-mediated transport
Ion channel transport
Iron uptake and transport
Miscellaneous transport and bi...
SLC-mediated transmembrane tra...
p.adjust
0.05
0.04
0.03
0.02
0.01
0.00

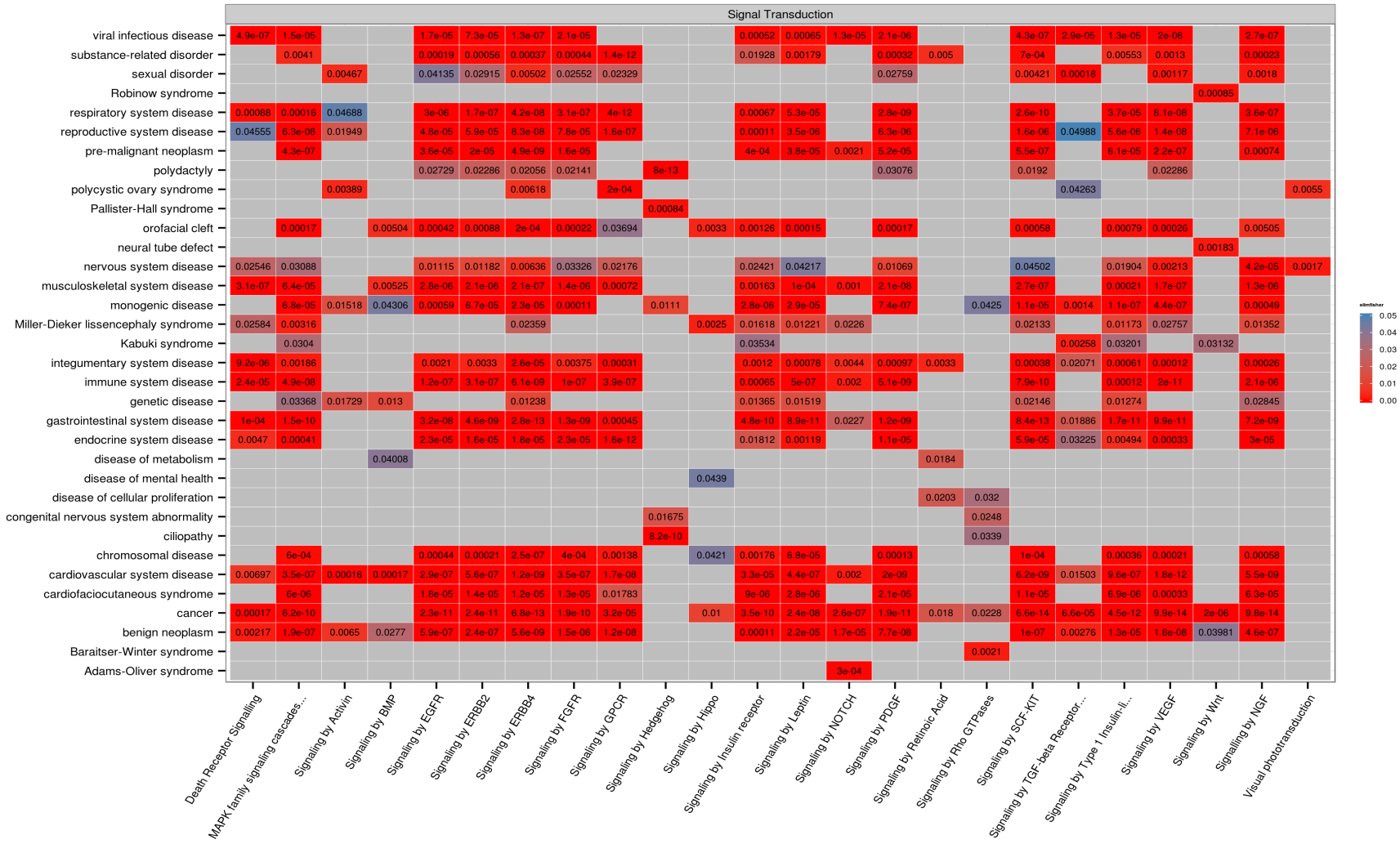
(f)


elementNum
0.05
0.04
0.03
0.02
0.01
0.00
Mitophagy
0.0046
0.0012
0.0317
Prrk_ Parkin Mediated Mitophagy...
viral infectious disease
nervous system disease
gastrointestinal system disease
DNA Repair
4.6e-08
3.4e-10
4.4e-05
0.006
0.00841
3.7e-07
0.013
0.00145
0.00508
0.003
4.1e-11
0.011
3e-04
0.003
0.0046
0.0264
0.0354
0.0043
0.00014
0.0081
0.0151
0.0117
0.003
0.00013
thoracic disease
neural tube defect
monogenic disease
inclusion body myopathy with Paget disea...
immune system disease
disease of cellular proliferation
congenital nervous system abnormality
chronic fatigue syndrome
cancer
Baraitser-Winter syndrome
Aicardi syndrome
viral infectious disease
Base Excision Repair
DNA Damage Bypass
DNA Double-Strand Break Repair...
Fanconi Anemia Pathway
Mismatch Repair
Nucleotide Excision Repair

(g)


Extracellular matrix organization
viral infectious disease
respiratory system disease
reproductive system disease
monogenic disease
gastrointestinal system disease
cardiovascular system disease
cancer
urinary system disease
musculoskeletal system disease
integumentary system disease
immune system disease
benign neoplasm
acrocallosal syndrome
Collagen formation
Degradation of the extracellul...
ECM proteoglycans
Elastic fibre formation
Integrin cell surface interact...
Laminin interactions
Non-integrin membrane-ECM inte...
Metabolism
visceral heterotaxy
VACTERL association
urinary system disease
syndrome
substance-related disorder
Proteus syndrome
polycystic ovary syndrome
orofacial cleft
Ollier disease
Ohdo syndrome
N syndrome
neural tube defect
musculoskeletal system disease
Miller-Dieker lissencephaly syndrome
maturity-onset diabetes of the young
Maffucci syndrome
integumentary system disease
inherited metabolic disorder
immune system disease
gastroschisis
Gamstorp-Wohlfart syndrome
endocrine system disease
disease by infectious agent
cognitive disorder
Coffin-Siris syndrome
chondrodysplasia punctata
benign neoplasm
Adams-Oliver syndrome
acrocallosal syndrome
acquired metabolic disease
Abacavir transport and metabol...
Biological oxidations
Inositol phosphate metabolism
Integration of energy metaboli...
Metabolism of amino acids and ...
Metabolism of carbohydrates
Metabolism of lipids and lipop...
Metabolism of nitric oxide
Metabolism of nucleotides
Metabolism of porphyrins
Metabolism of vitamins and cof...
Reversible hydration of carbon...
The citric acid (TCA) cycle an...
adjusted.p.value
0.05
0.04
0.03
0.02
0.01
0.00

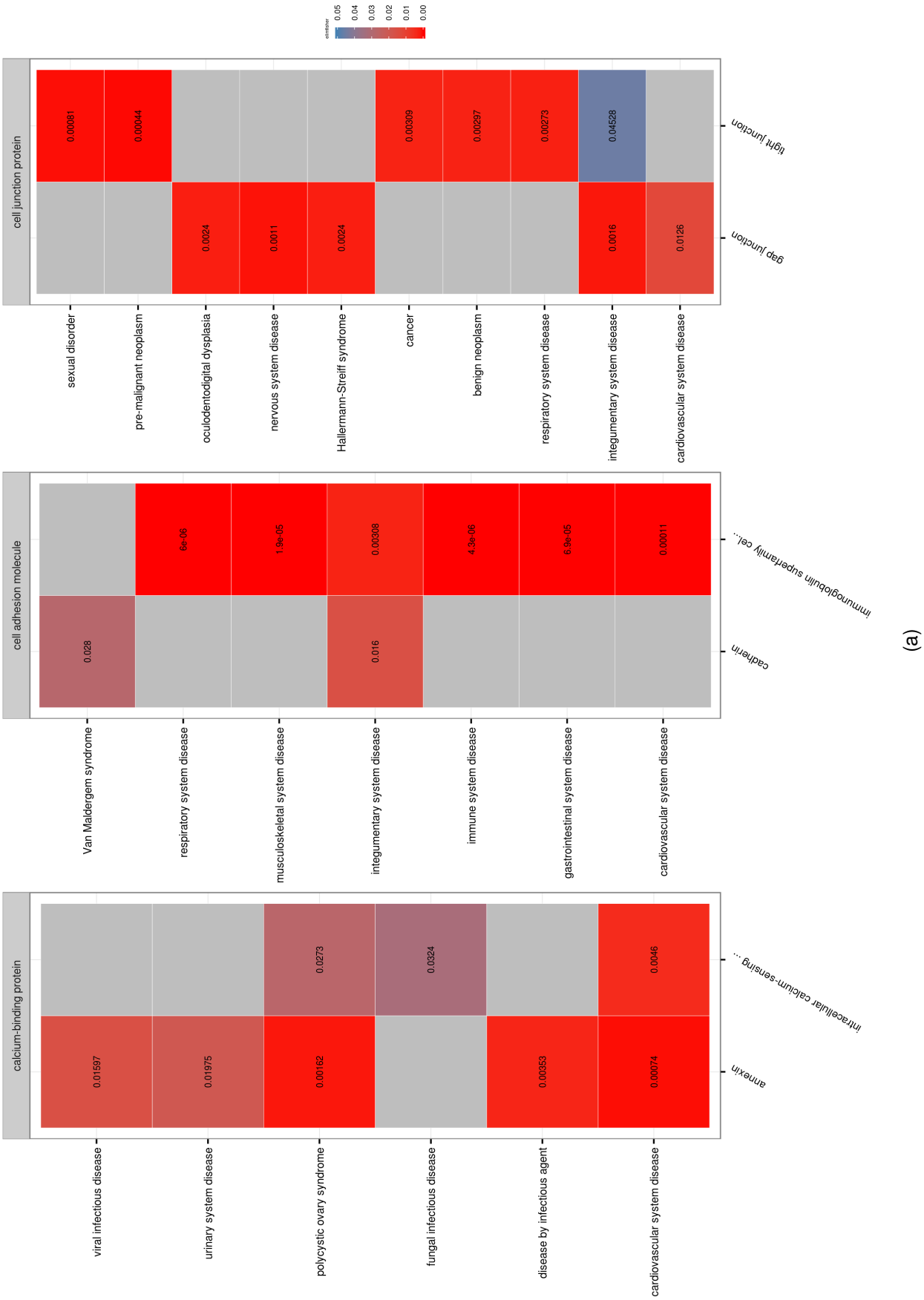
(h)



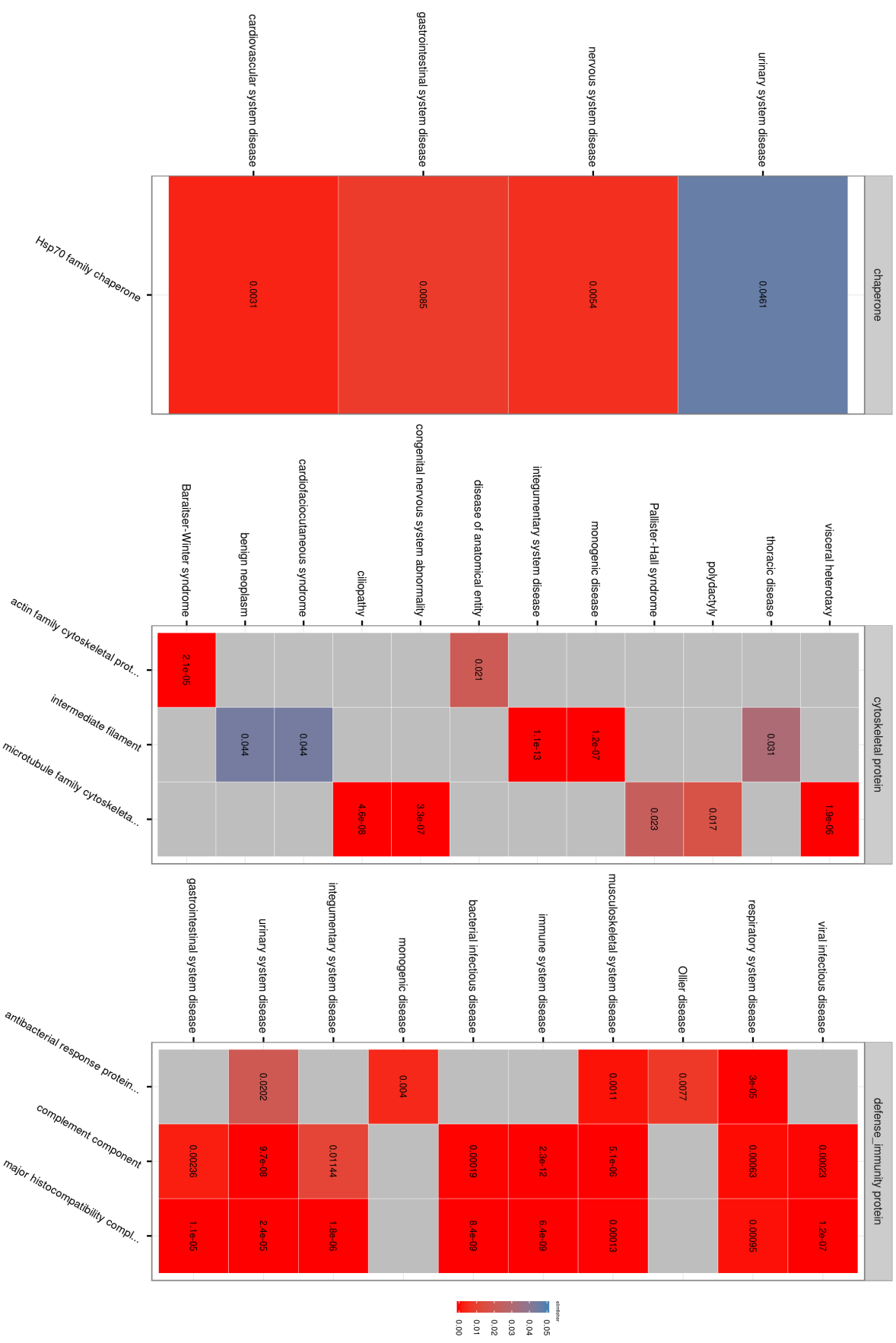
(i)

Figure A2: Heat map of disease profile of 128 level 2 Reactome pathways grouped by 23 top level Reactome pathways. Disease enrichment analysis were perform using *topOnto* and *hdgdb* with the *elim* topology methods. 136 level 3 HDO terms were selected to represent human disease and the top 5 enriched diseases for each RPO term were shown with the corresponding p-value.

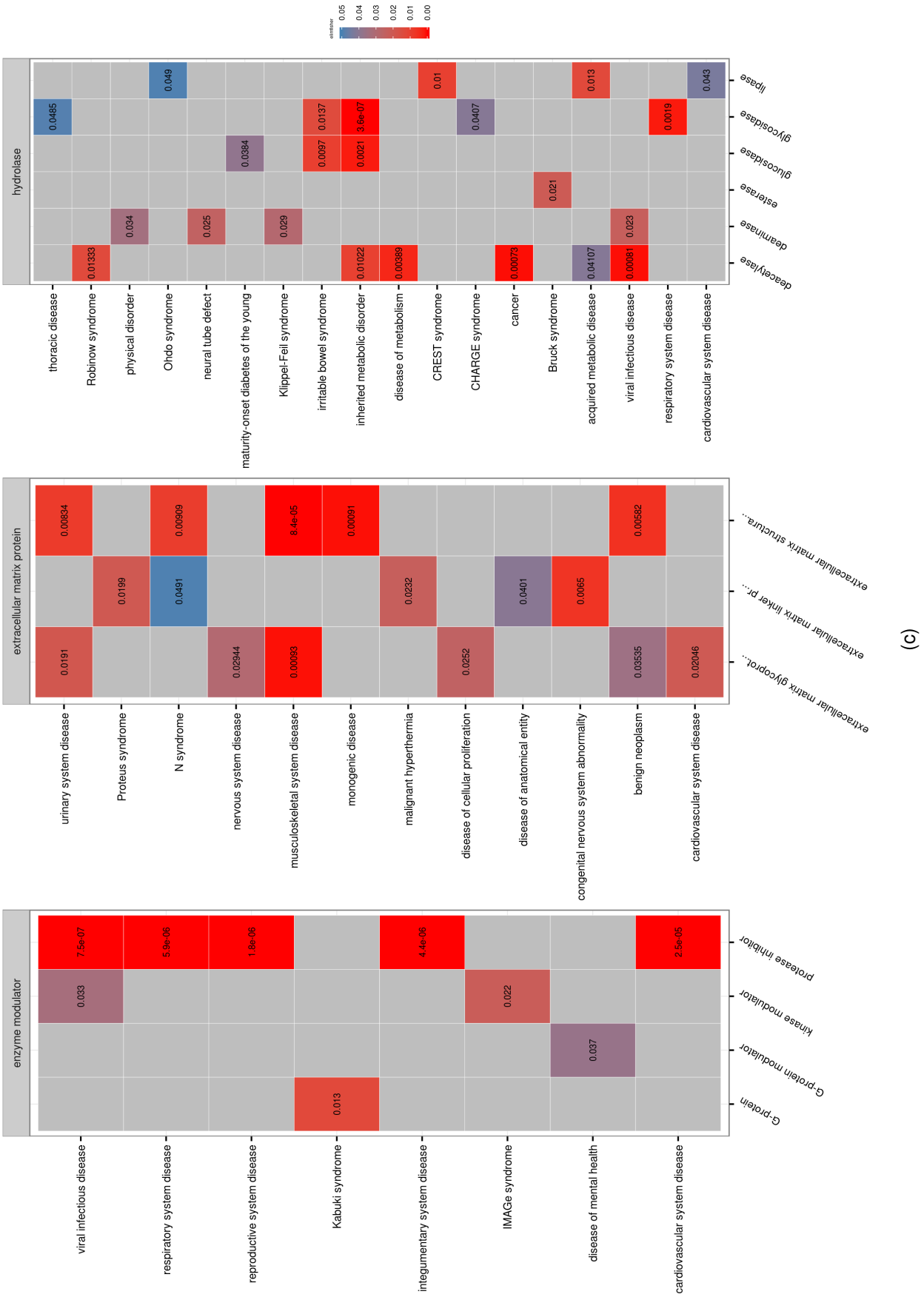
A.3 Panther protein class profiles of disease


cell junction protein
gap junction
tight junction
0.0016
0.0126
0.0024
0.0011
0.0024
0.00309
0.00297
0.00273
0.04528
0.00081
0.00044
sexual disorder
pre-malignant neoplasm
oculodentodigital dysplasia
nervous system disease
Hallermann-Streiff syndrome
cancer
benign neoplasm
respiratory system disease
integumentary system disease
cardiovascular system disease
cell adhesion molecule
cadherin
immunoglobulin superfamily cel...
0.016
0.028
6e-06
1.9e-05
0.00308
4.3e-06
6.9e-05
0.00011
Van Maldergem syndrome
respiratory system disease
musculoskeletal system disease
integumentary system disease
immune system disease
gastrointestinal system disease
cardiovascular system disease
calcium-binding protein
annexin
intracellular calcium-sensing ...
0.00074
0.00353
0.00162
0.01975
0.01597
0.0046
0.0324
0.0273
viral infectious disease
urinary system disease
polycystic ovary syndrome
fungal infectious disease
disease by infectious agent
cardiovascular system disease
elimFdr
0.05
0.04
0.03
0.02
0.01
0.00

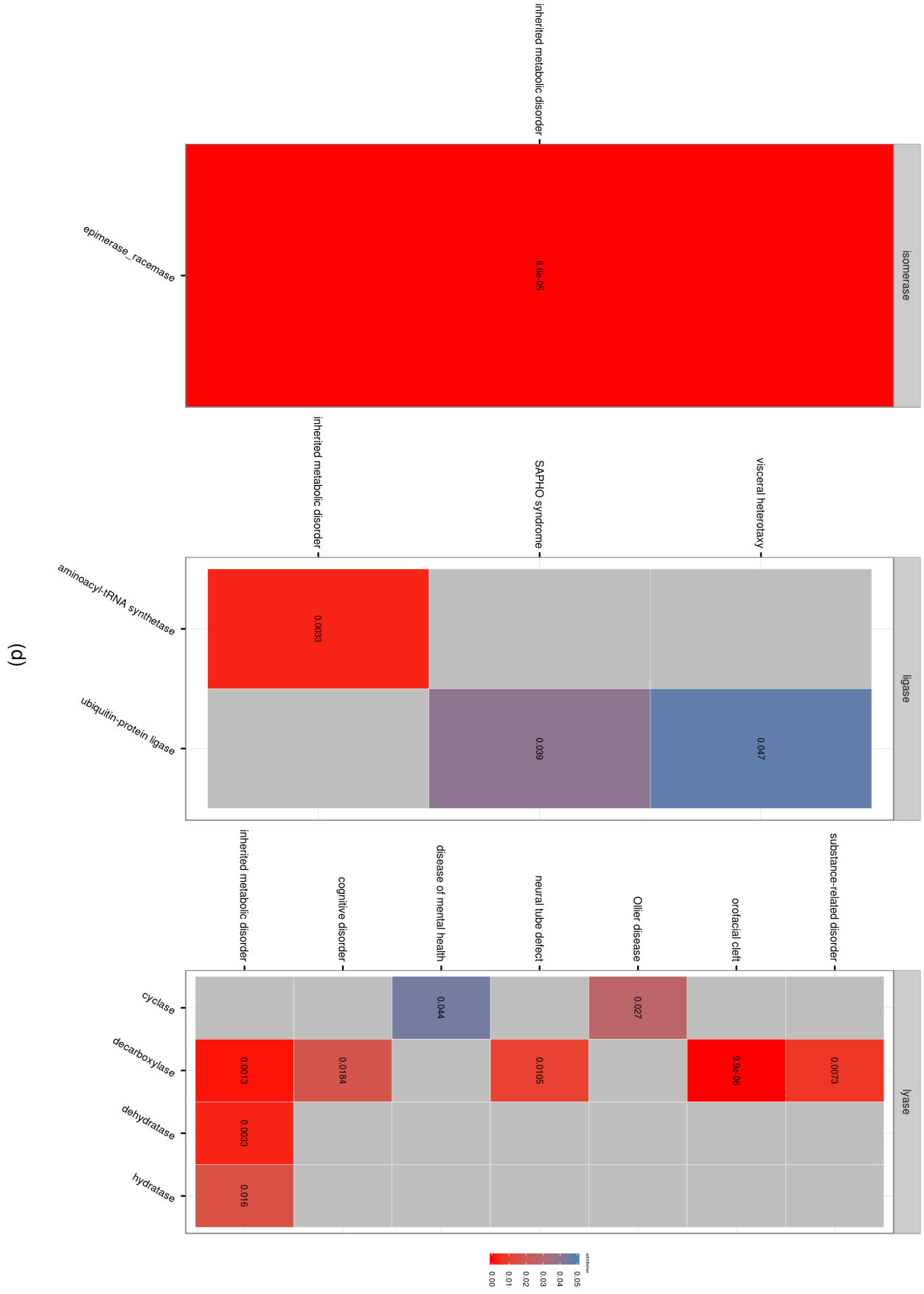
(a)

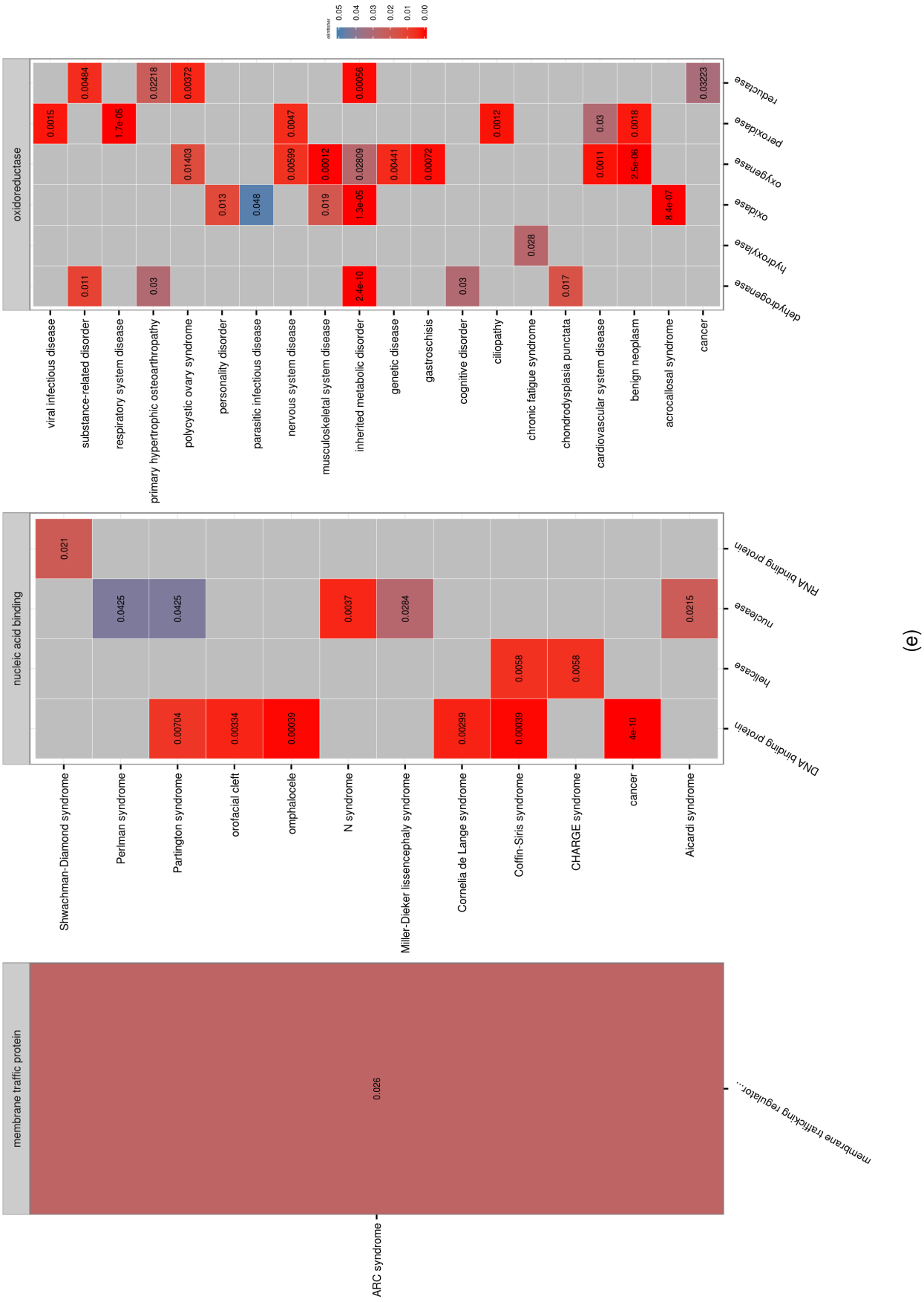

chaperone
urinary system disease
nervous system disease
gastrointestinal system disease
cardiovascular system disease
0.0461
0.0054
0.0085
0.0031
Hsp70 family chaperone
cytoskeletal protein
visceral heterotaxy
thoracic disease
polydactyly
Pallister-Hall syndrome
monogenic disease
integumentary system disease
disease of anatomical entity
congenital nervous system abnormality
ciliopathy
cardiofaciocutaneous syndrome
benign neoplasm
Baraitser-Winter syndrome
1.9e-06
0.031
0.017
0.023
1.2e-07
1.1e-13
0.021
3.3e-07
4.6e-08
0.044
0.044
2.1e-05
actin family cytoskeletal prot...
intermediate filament
microtubule family cytoskeleta...
defense_immunity protein
viral infectious disease
respiratory system disease
Ollier disease
musculoskeletal system disease
immune system disease
bacterial infectious disease
monogenic disease
integumentary system disease
urinary system disease
gastrointestinal system disease
0.00023
3e-05
0.0077
0.0011
2.3e-12
0.00019
0.004
0.01144
0.0202
0.00236
1.2e-07
0.00095
0.00013
6.4e-09
8.4e-09
1.8e-06
2.4e-05
1.1e-05
5.1e-06
0.00063
9.7e-08
antibacterial response protein...
complement component
major histocompatibility compl...
0.05
0.04
0.03
0.02
0.01
0.00

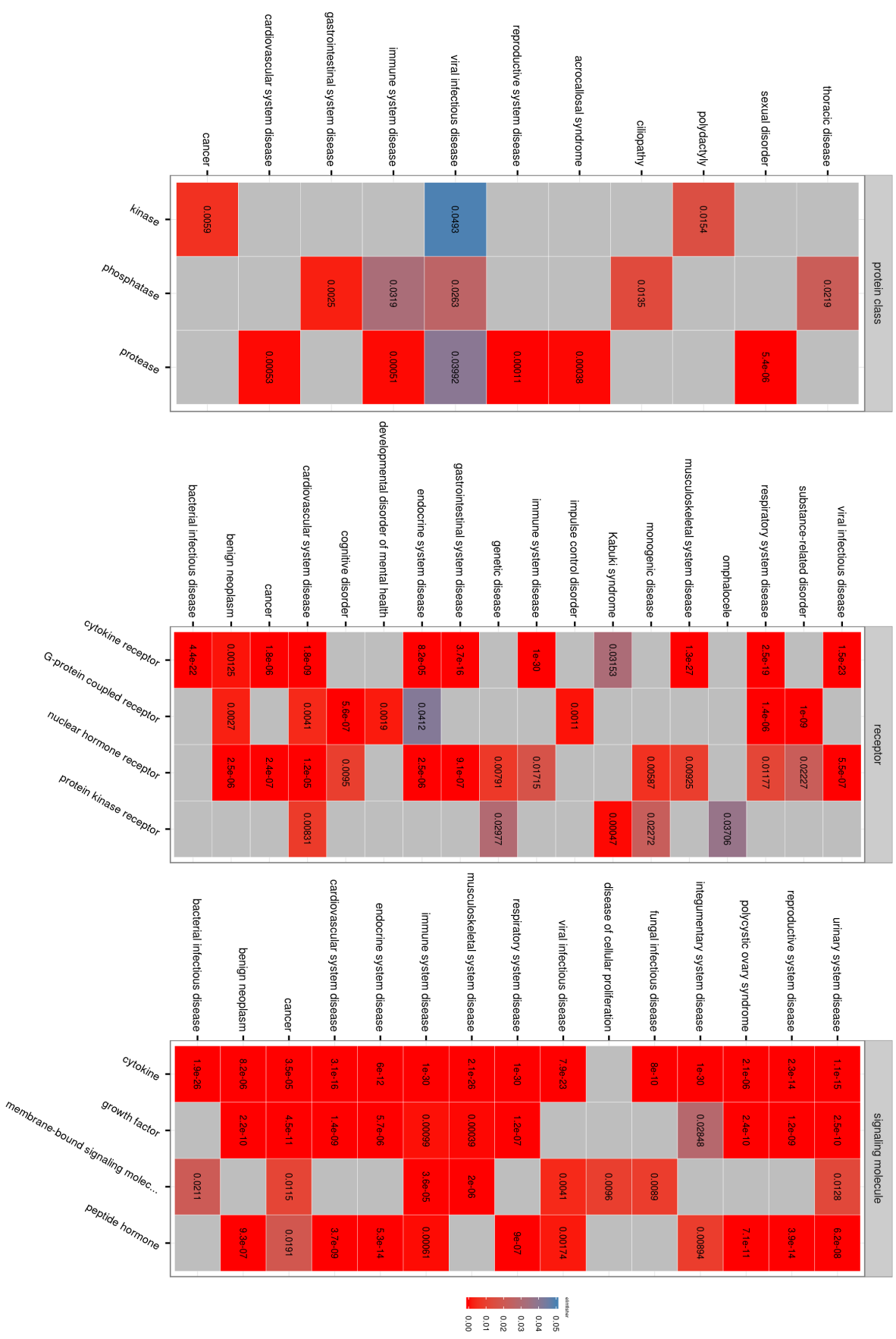
(b)


hydrolase
thoracic disease
Robinow syndrome
physical disorder
Ohdo syndrome
neural tube defect
maturity-onset diabetes of the young
Klippel-Feil syndrome
irritable bowel syndrome
inherited metabolic disorder
disease of metabolism
CREST syndrome
CHARGE syndrome
cancer
Bruck syndrome
acquired metabolic disease
viral infectious disease
respiratory system disease
cardiovascular system disease
deaceylase
deaminase
esterase
glucosidase
glycosidase
lipase
extracellular matrix protein
urinary system disease
Proteus syndrome
N syndrome
nervous system disease
musculoskeletal system disease
monogenic disease
malignant hyperthermia
disease of cellular proliferation
disease of anatomical entity
congenital nervous system abnormality
benign neoplasm
cardiovascular system disease
extracellular matrix glycoprot...
extracellular matrix linker pr...
extracellular matrix structura...
enzyme modulator
viral infectious disease
respiratory system disease
reproductive system disease
Kabuki syndrome
integumentary system disease
IMAGe syndrome
disease of mental health
cardiovascular system disease
G-protein
G-protein modulator
kinase modulator
protease inhibitor

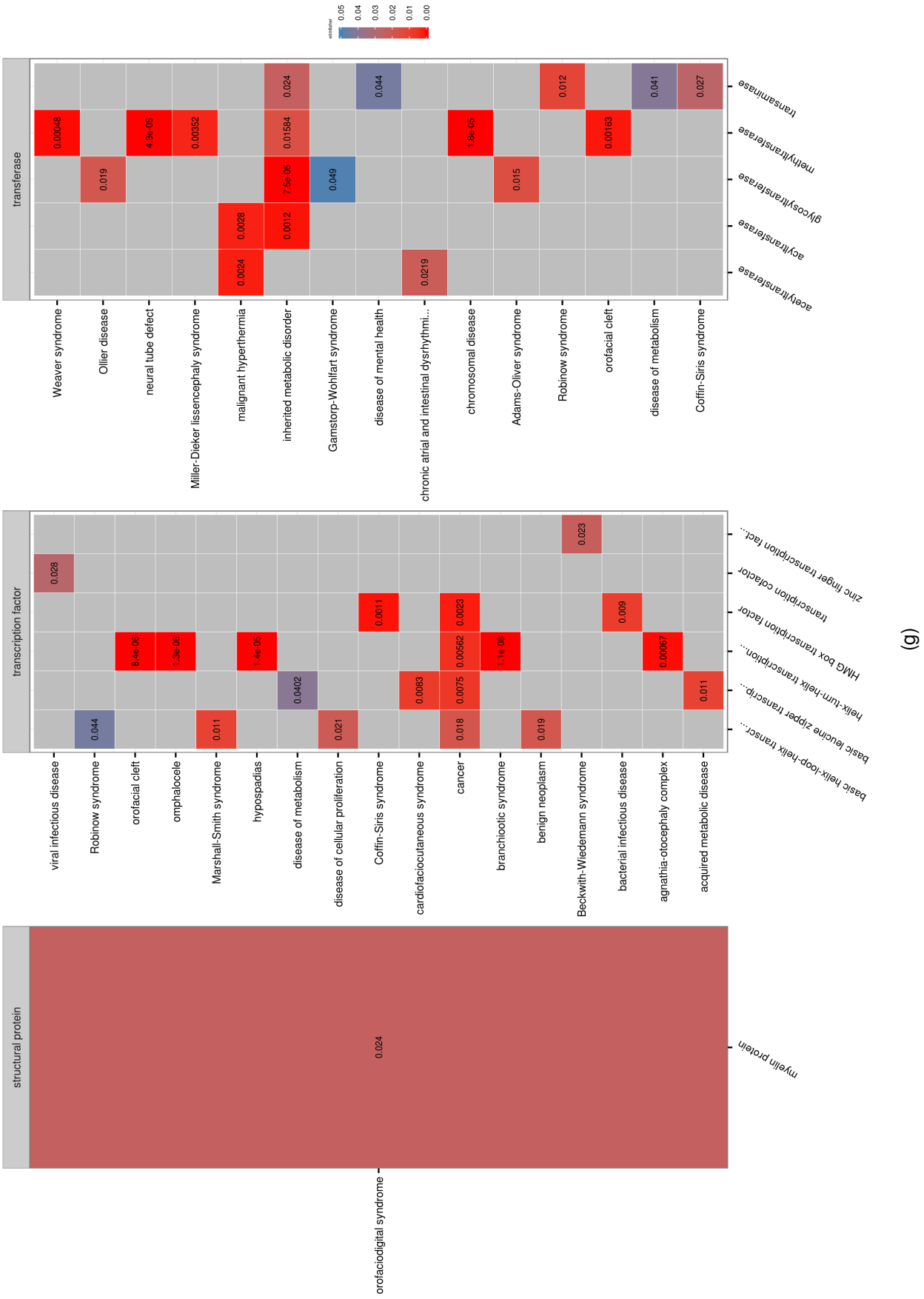
(c)


(d)
isomerase
ligase
lyase
epimerase_racemase
aminoacyl-tRNA synthetase
ubiquitin-protein ligase
cyclase
decarboxylase
dehydratase
hydratase
inherited metabolic disorder
SAPHO syndrome
visceral heterotaxy
cognitive disorder
disease of mental health
neural tube defect
Oliver disease
orofacial cleft
substance-related disorder
8.6e-05
0.0033
0.039
0.047
0.0013
0.0033
0.016
0.0184
0.044
0.0105
0.027
9.9e-06
0.0073
estimate
0.00
0.01
0.02
0.03
0.04
0.05

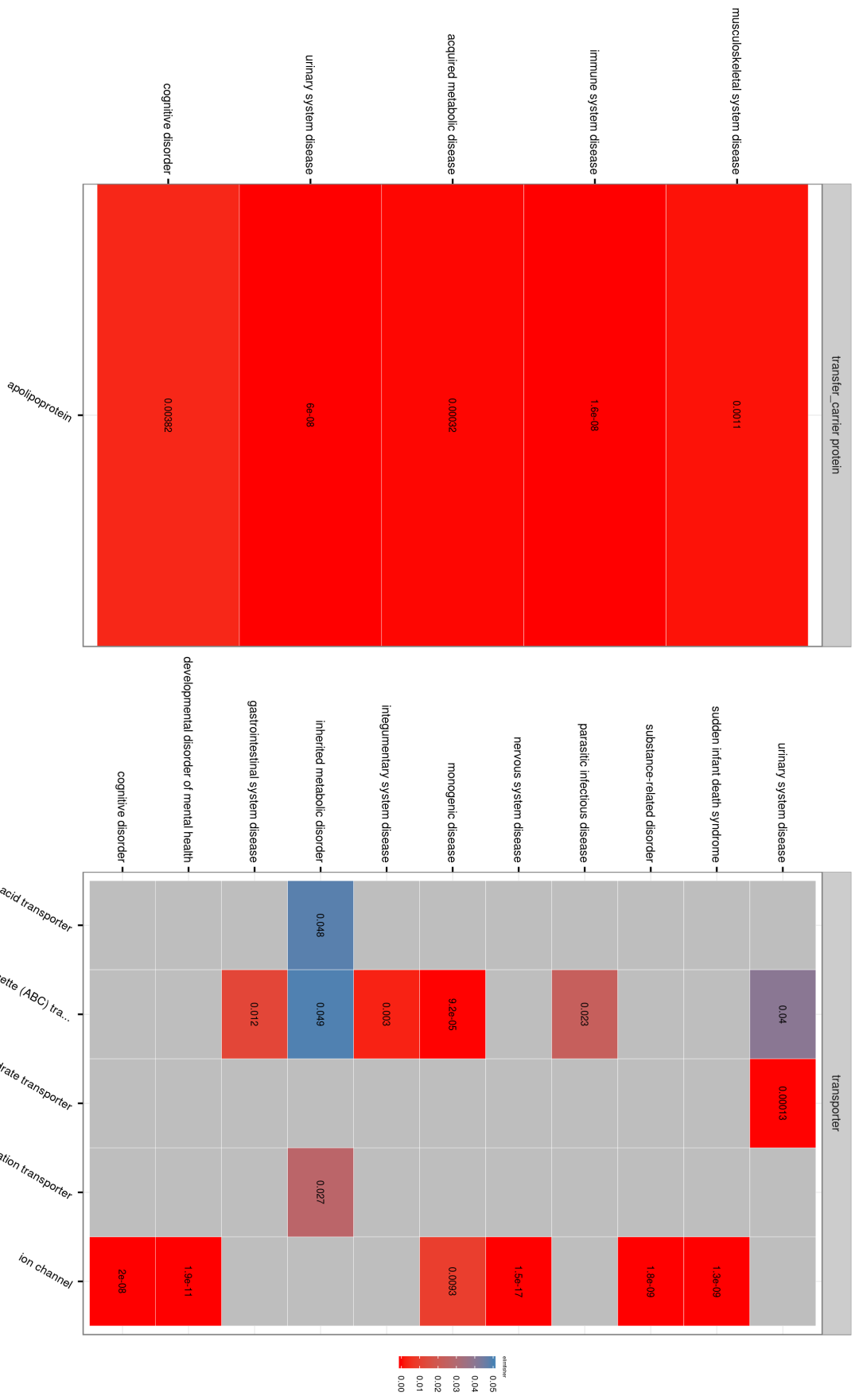

oxidoreductase
viral infectious disease
substance-related disorder
respiratory system disease
primary hypertrophic osteoarthropathy
polycystic ovary syndrome
personality disorder
parasitic infectious disease
nervous system disease
musculoskeletal system disease
inherited metabolic disorder
genetic disease
gastroschisis
cognitive disorder
ciliopathy
chronic fatigue syndrome
chondrodysplasia punctata
cardiovascular system disease
benign neoplasm
acrocallosal syndrome
cancer
dehydrogenase
hydroxylase
oxidase
oxygenase
peroxidase
reductase
nucleic acid binding
Shwachman-Diamond syndrome
Perlman syndrome
Partington syndrome
orofacial cleft
omphalocele
N syndrome
Miller-Dieker lissencephaly syndrome
Cornelia de Lange syndrome
Coffin-Siris syndrome
CHARGE syndrome
cancer
Aicardi syndrome
DNA binding protein
helicase
nuclease
RNA binding protein
membrane traffic protein
ARC syndrome
membrane trafficking regulator...
(e)


protein class
thoracic disease
sexual disorder
polydactyly
ciliopathy
acrocallosal syndrome
reproductive system disease
viral infectious disease
immune system disease
gastrointestinal system disease
cardiovascular system disease
cancer
kinase
phosphatase
protease
receptor
viral infectious disease
substance-related disorder
respiratory system disease
omphalocele
musculoskeletal system disease
monogenic disease
Kabuki syndrome
impulse control disorder
immune system disease
genetic disease
gastrointestinal system disease
endocrine system disease
developmental disorder of mental health
cognitive disorder
cardiovascular system disease
cancer
benign neoplasm
bacterial infectious disease
cytokine receptor
G-protein coupled receptor
nuclear hormone receptor
protein kinase receptor
signaling molecule
urinary system disease
reproductive system disease
polycystic ovary syndrome
integumentary system disease
fungal infectious disease
disease of cellular proliferation
viral infectious disease
respiratory system disease
musculoskeletal system disease
immune system disease
endocrine system disease
cardiovascular system disease
cancer
benign neoplasm
bacterial infectious disease
cytokine
growth factor
membrane-bound signaling molec...
peptide hormone
simdata
0.05
0.04
0.03
0.02
0.01
0.00

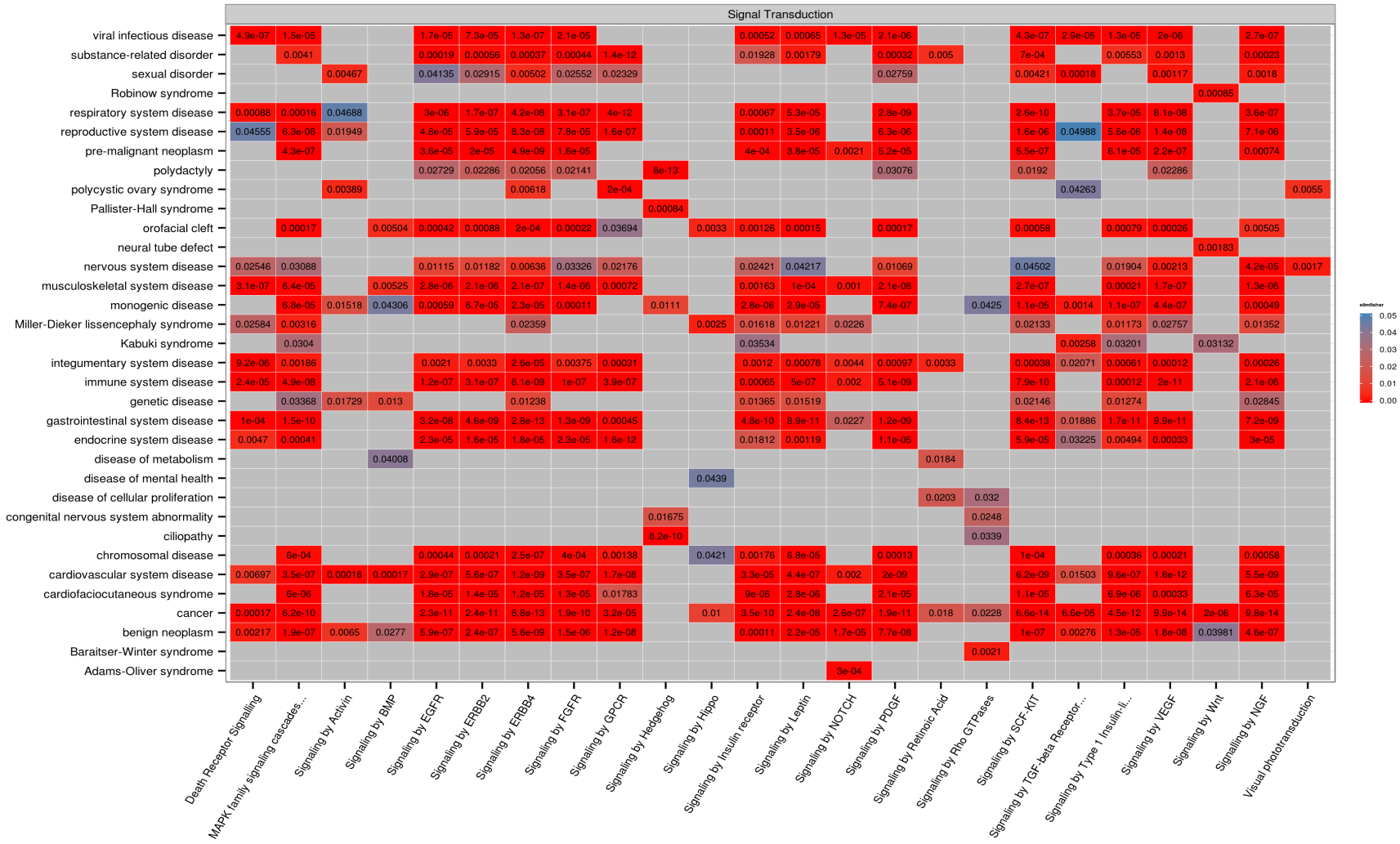
(f)


structural protein
transcription factor
transferase
myelin protein
basic helix-loop-helix transcr...
basic leucine zipper transcrip...
helix-turn-helix transcription...
HMG box transcription factor
transcription cofactor
zinc finger transcription fact...
acetyltransferase
acyltransferase
glycosyltransferase
methyltransferase
transaminase
orofaciodigital syndrome
viral infectious disease
Robinow syndrome
orofacial cleft
omphalocele
Marshall-Smith syndrome
hypospadias
disease of metabolism
disease of cellular proliferation
Coffin-Siris syndrome
cardiofaciocutaneous syndrome
cancer
branchiootic syndrome
benign neoplasm
Beckwith-Wiedemann syndrome
bacterial infectious disease
agnathia-otocephaly complex
acquired metabolic disease
Weaver syndrome
Ollier disease
neural tube defect
Miller-Dieker lissencephaly syndrome
malignant hyperthermia
inherited metabolic disorder
Gamstorp-Wohlfart syndrome
disease of mental health
chronic atrial and intestinal dysrhythmi...
chromosomal disease
Adams-Oliver syndrome
Robinow syndrome
orofacial cleft
disease of metabolism
Coffin-Siris syndrome
estimate
0.05
0.04
0.03
0.02
0.01
0.00

(g)


transfer_carrier protein
musculoskeletal system disease
immune system disease
acquired metabolic disease
urinary system disease
cognitive disorder
0.0011
1.6e-08
0.00032
6e-08
0.00392
apolipoprotein
transporter
urinary system disease
sudden infant death syndrome
substance-related disorder
parasitic infectious disease
nervous system disease
monogenic disease
integumentary system disease
inherited metabolic disorder
gastrointestinal system disease
developmental disorder of mental health
cognitive disorder
0.04
0.00013
1.3e-09
1.8e-09
0.023
1.5e-17
9.2e-05
0.0093
0.003
0.048
0.049
0.012
0.027
2e-08
1.9e-11
amino acid transporter
ATP-binding cassette (ABC) tra...
carbohydrate transporter
cation transporter
ion channel
0.05
0.04
0.03
0.02
0.01
0.00

(h)



(i)

Figure A3: Heat map of disease profile of 73 level 3 protein class grouped by 23 level 2 PCO terms. Disease enrichment analysis were perform using *topOnto* and *hdgdb* with the *elim* topology methods. 136 level 3 HDO terms were selected to represent human disease and the top 5 enriched diseases for each PCO term were shown with the corresponding p-value.

A.4 Enrichment analysis results of the ARC complex

TERM.ID	Term	Level	classic	elim	weight01	parentchild
HP:0100753	Schizophrenia	7	7.7e-20	7.7e-20	7.7e-20	2.9e-07
HP:0012638	Abnormality of nervous system physiology	5	7.0e-16	0.42183	1.00000	9.4e-07
HP:0000708	Behavioral abnormality	6	7.1e-15	0.67728	1.00000	0.00426
HP:0001250	Seizures	6	7.1e-12	4.4e-09	2.5e-08	5.2e-06
HP:0000707	Abnormality of the nervous system	4	1.8e-10	0.80585	1.00000	1.9e-10
HP:0012639	Abnormality of nervous system morphology	5	2.4e-07	0.17432	1.00000	0.09045
HP:0002011	Morphological abnormality of the central...	6	5.4e-07	0.15205	1.00000	0.53314
HP:0001298	Encephalopathy	6	6.0e-07	0.00327	0.01815	0.00028
HP:0002529	Neuronal loss in central nervous system	8	6.3e-06	6.3e-06	6.3e-06	0.00019
HP:0012823	Clinical modifier	3	2.0e-05	0.01382	1.00000	2.0e-05

(a) classic

TERM.ID	Term	Level	classic	elim	weight01	parentchild
HP:0100753	Schizophrenia	7	7.7e-20	7.7e-20	7.7e-20	2.9e-07
HP:0001250	Seizures	6	7.1e-12	4.4e-09	2.5e-08	5.2e-06
HP:0002529	Neuronal loss in central nervous system	8	6.3e-06	6.3e-06	6.3e-06	0.00019
HP:0200134	Epileptic encephalopathy	7	3.7e-05	3.7e-05	3.7e-05	0.34006
HP:0003745	Sporadic	4	7.0e-05	7.0e-05	7.0e-05	0.01010
HP:0000726	Dementia	9	0.00015	0.00015	0.00058	0.01383
HP:0002511	Alzheimer disease	7	0.00021	0.00021	0.00021	0.09168
HP:0007105	Infantile encephalopathy	7	0.00045	0.00045	0.00045	0.29452
HP:0002539	Cortical dysplasia	11	0.00057	0.00057	0.00057	0.32578
HP:0000733	Stereotypic behavior	7	0.00070	0.00070	0.00070	0.00442

(b) elim

TERM.ID	Term	Level	classic	elim	weight01	parentchild
HP:0100753	Schizophrenia	7	7.7e-20	7.7e-20	7.7e-20	2.9e-07
HP:0001250	Seizures	6	7.1e-12	4.4e-09	2.5e-08	5.2e-06
HP:0002529	Neuronal loss in central nervous system	8	6.3e-06	6.3e-06	6.3e-06	0.00019
HP:0200134	Epileptic encephalopathy	7	3.7e-05	3.7e-05	3.7e-05	0.34006
HP:0003745	Sporadic	4	7.0e-05	7.0e-05	7.0e-05	0.01010
HP:0002511	Alzheimer disease	7	0.00021	0.00021	0.00021	0.09168
HP:0007105	Infantile encephalopathy	7	0.00045	0.00045	0.00045	0.29452
HP:0002539	Cortical dysplasia	11	0.00057	0.00057	0.00057	0.32578
HP:0000726	Dementia	9	0.00015	0.00015	0.00058	0.01383
HP:0000733	Stereotypic behavior	7	0.00070	0.00070	0.00070	0.00442

(c) weight01

TERM.ID	Term	Level	classic	elim	weight01	parentchild
HP:0000707	Abnormality of the nervous system	4	1.8e-10	0.80585	1.00000	1.9e-10
HP:0100753	Schizophrenia	7	7.7e-20	7.7e-20	7.7e-20	2.9e-07
HP:0012638	Abnormality of nervous system physiology	5	7.0e-16	0.42183	1.00000	9.4e-07
HP:0001250	Seizures	6	7.1e-12	4.4e-09	2.5e-08	5.2e-06
HP:0012823	Clinical modifier	3	2.0e-05	0.01382	1.00000	2.0e-05
HP:0000005	Mode of inheritance	3	7.5e-05	0.01812	0.32023	7.5e-05
HP:0002529	Neuronal loss in central nervous system	8	6.3e-06	6.3e-06	6.3e-06	0.00019
HP:0001298	Encephalopathy	6	6.0e-07	0.00327	0.01815	0.00028
HP:0100547	Abnormality of forebrain morphology	8	5.4e-05	1.00000	1.00000	0.00339
HP:0011355	Localized skin lesion	6	0.00592	0.28545	1.00000	0.00373

(d) parentchild

Table A1: ARC complexes HPO enrichment result

TERM.ID	Term	Level	classic	elim	weight01	parentchild
R-HSA-112315	Transmission across Chemical Synapses	3	4.2e-18	0.05137	1.00000	0.00458
R-HSA-112316	Neuronal System	2	3.2e-16	0.04956	1.00000	3.2e-16
R-HSA-438066	Unblocking of NMDA receptor, glutamate...	6	7.5e-13	7.5e-13	7.5e-13	0.00264
R-HSA-442755	Activation of NMDA receptor upon glutama...	5	4.6e-11	0.13060	1.00000	0.00137
R-HSA-112314	Neurotransmitter Receptor Binding And Do...	4	3.0e-10	0.03488	1.00000	0.86560
R-HSA-442729	CREB phosphorylation through the activat...	7	2.2e-09	2.2e-09	2.2e-09	0.01584
R-HSA-442742	CREB phosphorylation through the activat...	7	3.0e-09	0.08115	1.00000	0.13206
R-HSA-442982	Ras activation uopn Ca2+ infux through N...	8	5.4e-09	5.4e-09	5.4e-09	0.16126
R-HSA-399719	Trafficking of AMPA receptors	6	6.8e-09	5.1e-06	5.1e-06	1.00000
R-HSA-399721	Glutamate Binding, Activation of AMPA Re...	5	6.8e-09	1.00000	1.00000	0.00720

(a) classic

TERM.ID	Term	Level	classic	elim	weight01	parentchild
R-HSA-438066	Unblocking of NMDA receptor, glutamate...	6	7.5e-13	7.5e-13	7.5e-13	0.00264
R-HSA-442729	CREB phosphorylation through the activat...	7	2.2e-09	2.2e-09	2.2e-09	0.01584
R-HSA-442982	Ras activation uopn Ca2+ infux through N...	8	5.4e-09	5.4e-09	5.4e-09	0.16126
R-HSA-5578775	Ion homeostasis	4	1.5e-08	1.5e-08	1.5e-08	0.00181
R-HSA-936837	Ion transport by P-type ATPases	4	2.8e-08	2.8e-08	2.8e-08	0.00753
R-HSA-5682910	LGI-ADAM interactions	3	1.0e-07	1.0e-07	1.0e-07	1.8e-05
R-HSA-399719	Trafficking of AMPA receptors	6	6.8e-09	5.1e-06	5.1e-06	1.00000
R-HSA-888590	GABA synthesis, release, reuptake and de...	5	1.2e-08	7.3e-06	0.00017	0.02655
R-HSA-210500	Glutamate Neurotransmitter Release Cycle	5	5.9e-05	5.9e-05	5.9e-05	0.60134
R-HSA-389957	Prefoldin mediated transfer of substrate...	6	8.2e-05	8.2e-05	8.2e-05	0.47513

(b) elim

TERM.ID	Term	Level	classic	elim	weight01	parentchild
R-HSA-438066	Unblocking of NMDA receptor, glutamate b...	6	7.5e-13	7.5e-13	7.5e-13	0.00264
R-HSA-442729	CREB phosphorylation through the activat...	7	2.2e-09	2.2e-09	2.2e-09	0.01584
R-HSA-442982	Ras activation uopn Ca2+ infux through N...	8	5.4e-09	5.4e-09	5.4e-09	0.16126
R-HSA-5578775	Ion homeostasis	4	1.5e-08	1.5e-08	1.5e-08	0.00181
R-HSA-936837	Ion transport by P-type ATPases	4	2.8e-08	2.8e-08	2.8e-08	0.00753
R-HSA-5682910	LGI-ADAM interactions	3	1.0e-07	1.0e-07	1.0e-07	1.8e-05
R-HSA-399719	Trafficking of AMPA receptors	6	6.8e-09	5.1e-06	5.1e-06	1.00000
R-HSA-210500	Glutamate Neurotransmitter Release Cycle	5	5.9e-05	5.9e-05	5.9e-05	0.60134
R-HSA-389957	Prefoldin mediated transfer of substrate...	6	8.2e-05	8.2e-05	8.2e-05	0.47513
R-HSA-77387	Insulin receptor recycling	4	8.2e-05	8.2e-05	8.2e-05	0.01819

(c) weight01

TERM.ID	Term	Level	classic	elim	weight01	parentchild
R-HSA-112316	Neuronal System	2	3.2e-16	0.04956	1.00000	3.2e-16
R-HSA-382551	TM transport of small molecul...	2	2.0e-07	0.02895	1.0000	2.0e-07
R-HSA-1266738	Developmental Biology	2	1.1e-06	0.44056	1.00000	1.1e-06
R-HSA-74752	Signaling by Insulin receptor	3	2.8e-06	1.00000	1.00000	1.2e-06
R-HSA-5682910	LGI-ADAM interactions	3	1.0e-07	1.0e-07	1.0e-07	1.8e-05
R-HSA-397014	Muscle contraction	2	6.7e-05	0.70095	1.00000	6.7e-05
R-HSA-2586552	Signaling by Leptin	3	0.00026	1.00000	1.00000	0.00026
R-HSA-5673001	RAF/MAP kinase cascade	8	0.00020	0.00020	0.00028	0.00043
R-HSA-194138	Signaling by VEGF	3	0.00059	0.54018	1.00000	0.00056
R-HSA-177929	Signaling by EGFR	3	0.00091	0.60478	1.00000	0.00087

(d) parentchild

Table A2: ARC complexes RECTOMEPATHWAY enrichment result

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0007268	synaptic transmission	9	1.4e-28	3.0e-10	7.9e-09	1.00000
GO:0099536	synaptic signaling	7	1.4e-28	1.00000	1.00000	5.0e-07
GO:0099537	trans-synaptic signaling	8	1.4e-28	1.00000	1.00000	1.00000
GO:0007267	cell-cell signaling	6	5.2e-23	0.38090	0.25777	7.1e-16
GO:0006811	ion transport	7	8.5e-21	0.38187	1.00000	9.4e-09
GO:0050804	modulation of synaptic transmission	10	8.6e-16	0.00010	0.06192	2.6e-09
GO:0030001	metal ion transport	9	6.2e-15	0.14213	1.00000	0.04726
GO:0034220	ion transmembrane transport	8	7.8e-15	0.00465	0.00014	0.04620
GO:0006812	cation transport	8	1.3e-14	0.08283	1.00000	0.46859
GO:0055085	transmembrane transport	7	2.2e-14	0.04115	0.03305	3.3e-11

(a) classic

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0007268	synaptic transmission	9	1.4e-28	3.0e-10	7.9e-09	1.00000
GO:0035235	ionotropic glutamate receptor signaling ...	9	7.0e-09	7.0e-09	7.0e-09	0.02103
GO:0035249	synaptic transmission, glutamatergic	11	2.4e-08	2.4e-08	3.3e-08	0.39242
GO:0015991	ATP hydrolysis coupled proton transport	13	2.4e-08	2.4e-08	2.4e-08	1.00000
GO:0036376	sodium ion export from cell	13	5.3e-08	5.3e-08	2.4e-07	0.27778
GO:0030007	cellular potassium ion homeostasis	11	1.1e-07	1.1e-07	1.1e-07	2.1e-05
GO:0006883	cellular sodium ion homeostasis	11	7.3e-07	7.3e-07	7.3e-07	0.00013
GO:0010107	potassium ion import	13	1.1e-06	1.1e-06	2.8e-05	0.00052
GO:0033572	transferrin transport	13	1.9e-06	1.9e-06	1.9e-06	2.1e-05
GO:0007612	learning	8	1.3e-10	4.1e-06	4.8e-07	0.01308

(b) elim

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0035235	ionotropic glutamate receptor signaling ...	9	7.0e-09	7.0e-09	7.0e-09	0.02103
GO:0007268	synaptic transmission	9	1.4e-28	3.0e-10	7.9e-09	1.00000
GO:0015991	ATP hydrolysis coupled proton transport	13	2.4e-08	2.4e-08	2.4e-08	1.00000
GO:0035249	synaptic transmission, glutamatergic	11	2.4e-08	2.4e-08	3.3e-08	0.39242
GO:0030007	cellular potassium ion homeostasis	11	1.1e-07	1.1e-07	1.1e-07	2.1e-05
GO:0036376	sodium ion export from cell	13	5.3e-08	5.3e-08	2.4e-07	0.27778
GO:0007612	learning	8	1.3e-10	4.1e-06	4.8e-07	0.01308
GO:0006883	cellular sodium ion homeostasis	11	7.3e-07	7.3e-07	7.3e-07	0.00013
GO:0033572	transferrin transport	13	1.9e-06	1.9e-06	1.9e-06	2.1e-05
GO:0086064	cell communication by electrical couplin...	11	4.4e-06	4.4e-06	4.4e-06	0.04663

(c) weight01

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0007267	cell-cell signaling	6	5.2e-23	0.38090	0.25777	7.1e-16
GO:0055085	transmembrane transport	7	2.2e-14	0.04115	0.03305	3.3e-11
GO:1902578	single-organism localization	4	5.1e-14	0.95988	1.00000	5.1e-11
GO:0051179	localization	3	7.3e-10	0.93114	1.00000	7.3e-10
GO:0050804	modulation of synaptic transmission	10	8.6e-16	0.00010	0.06192	2.6e-09
GO:0023052	signaling	3	3.1e-09	0.90943	1.00000	3.1e-09
GO:0006811	ion transport	7	8.5e-21	0.38187	1.00000	9.4e-09
GO:0044763	single-organism cellular process	4	7.2e-10	0.23259	1.00000	1.7e-08
GO:0007399	nervous system development	7	1.1e-09	0.00447	0.08974	2.3e-08
GO:0044708	single-organism behavior	4	1.1e-08	0.53844	1.00000	1.2e-07

(d) parentchild

Table A3: ARC complexes GOBP enrichment result

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0097458	neuron part	5	5.8e-18	0.01328	1.00000	9.2e-17
GO:0043005	neuron projection	6	1.3e-17	0.01162	0.00290	4.1e-05
GO:0071944	cell periphery	5	1.8e-16	0.19275	1.00000	1.6e-14
GO:0098805	whole membrane	4	2.7e-16	0.45342	1.00000	3.0e-10
GO:0042995	cell projection	5	8.6e-16	0.06371	0.37528	1.6e-14
GO:0098590	plasma membrane region	8	2.1e-15	1.00000	1.00000	2.9e-07
GO:0045202	synapse	3	3.8e-15	0.04574	0.01386	3.8e-15
GO:0098589	membrane region	5	4.4e-15	0.18066	1.00000	2.8e-10
GO:0005886	plasma membrane	6	7.3e-15	5.8e-06	1.1e-05	1.4e-13
GO:0031982	vesicle	4	8.1e-15	0.13391	0.00601	7.3e-14

(a) classic

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0045211	postsynaptic membrane	10	8.4e-12	8.4e-12	8.4e-12	0.16211
GO:0030054	cell junction	3	3.2e-13	4.3e-11	1.0e-14	3.2e-13
GO:0070062	extracellular exosome	7	4.9e-10	4.9e-10	4.9e-10	0.71396
GO:0030666	endocytic vesicle membrane	12	4.6e-09	4.6e-09	7.4e-08	2.1e-05
GO:0097481	neuronal postsynaptic density	7	5.5e-08	5.5e-08	5.5e-08	0.00011
GO:0008021	synaptic vesicle	11	3.2e-07	3.2e-07	2.4e-06	0.00651
GO:0005890	sodium:potassium-exchanging ATPase...	10	3.8e-07	3.8e-07	3.8e-07	2.1e-05
GO:0014069	postsynaptic density	6	8.7e-12	2.2e-06	2.2e-06	2.6e-11
GO:0005886	plasma membrane	6	7.3e-15	5.8e-06	1.1e-05	1.4e-13
GO:0043197	dendritic spine	8	8.6e-07	1.2e-05	1.0e-05	0.01363

(b) elim

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0030054	cell junction	3	3.2e-13	4.3e-11	1.0e-14	3.2e-13
GO:0045211	postsynaptic membrane	10	8.4e-12	8.4e-12	8.4e-12	0.16211
GO:0070062	extracellular exosome	7	4.9e-10	4.9e-10	4.9e-10	0.71396
GO:0097481	neuronal postsynaptic density	7	5.5e-08	5.5e-08	5.5e-08	0.00011
GO:0030666	endocytic vesicle membrane	12	4.6e-09	4.6e-09	7.4e-08	2.1e-05
GO:0005890	sodium:potassium-exchanging ATPase...	10	3.8e-07	3.8e-07	3.8e-07	2.1e-05
GO:0014069	postsynaptic density	6	8.7e-12	2.2e-06	2.2e-06	2.6e-11
GO:0008021	synaptic vesicle	11	3.2e-07	3.2e-07	2.4e-06	0.00651
GO:0043197	dendritic spine	8	8.6e-07	1.2e-05	1.0e-05	0.01363
GO:0005886	plasma membrane	6	7.3e-15	5.8e-06	1.1e-05	1.4e-13

(c) weight01

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0097458	neuron part	5	5.8e-18	0.01328	1.00000	9.2e-17
GO:0045202	synapse	3	3.8e-15	0.04574	0.01386	3.8e-15
GO:0044456	synapse part	4	1.5e-14	1.00000	1.00000	1.5e-14
GO:0042995	cell projection	5	8.6e-16	0.06371	0.37528	1.6e-14
GO:0071944	cell periphery	5	1.8e-16	0.19275	1.00000	1.6e-14
GO:0031982	vesicle	4	8.1e-15	0.13391	0.00601	7.3e-14
GO:0005886	plasma membrane	6	7.3e-15	5.8e-06	1.1e-05	1.4e-13
GO:0030054	cell junction	3	3.2e-13	4.3e-11	1.0e-14	3.2e-13
GO:0098794	postsynapse	5	6.6e-14	1.00000	1.00000	4.6e-13
GO:0031988	membrane-bounded vesicle	5	5.9e-14	0.49712	1.00000	6.4e-13

(d) parentchild

Table A4: ARC complexes GOCC enrichment result

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0022891	substrate-specific transmembrane transpo...	5	1.7e-15	0.49631	1.00000	0.00111
GO:0015075	ion transmembrane transporter activity	6	1.9e-15	0.75058	1.00000	0.33200
GO:0022890	inorganic cation transmembrane transport...	8	2.4e-15	1.00000	1.00000	0.00936
GO:0022857	transmembrane transporter activity	4	2.1e-14	0.60800	1.00000	0.00038
GO:0022892	substrate-specific transporter activity	4	8.3e-14	0.66654	1.00000	0.00192
GO:0008324	cation transmembrane transporter activit...	7	1.6e-13	1.00000	1.00000	0.29821
GO:0015077	monovalent inorganic cation transmembran...	9	3.2e-12	0.00026	1.00000	0.36195
GO:0022804	active transmembrane transporter activit...	5	5.2e-12	0.04140	1.00000	0.00571
GO:0019829	cation-transporting ATPase activity	14	6.5e-12	0.02009	1.00000	6.1e-05
GO:0042625	ATPase activity, coupled to transmembran...	13	9.0e-12	1.00000	1.00000	4.1e-05

(a) classic

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0005234	extracellular-glutamate-gated ion channe...	12	1.3e-09	1.3e-09	1.3e-09	0.00194
GO:0005391	sodium:potassium-exchanging ATPase activ...	16	9.8e-08	9.8e-08	9.8e-08	0.00118
GO:0023026	MHC class II protein complex binding	7	2.5e-06	2.5e-06	2.5e-06	1.00000
GO:0004972	NMDA glutamate receptor activity	12	1.3e-05	1.3e-05	1.3e-05	0.33591
GO:0019901	protein kinase binding	7	2.8e-05	2.8e-05	0.00055	0.29117
GO:0005200	structural constituent of cytoskeleton	4	3.2e-05	3.2e-05	3.2e-05	0.00891
GO:0042288	MHC class I protein binding	7	6.5e-05	6.5e-05	6.5e-05	0.18571
GO:0005516	calmodulin binding	5	8.7e-05	8.7e-05	8.7e-05	0.00091
GO:0046961	proton-transporting ATPase activity, rot...	16	0.00015	0.00015	0.00015	0.07025
GO:0032403	protein complex binding	5	1.6e-06	0.00019	0.00115	0.00021

(b) elim

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0005234	extracellular-glutamate-gated ion channe...	12	1.3e-09	1.3e-09	1.3e-09	0.00194
GO:0005391	sodium:potassium-exchanging ATPase activ...	16	9.8e-08	9.8e-08	9.8e-08	0.00118
GO:0023026	MHC class II protein complex binding	7	2.5e-06	2.5e-06	2.5e-06	1.00000
GO:0004972	NMDA glutamate receptor activity	12	1.3e-05	1.3e-05	1.3e-05	0.33591
GO:0005200	structural constituent of cytoskeleton	4	3.2e-05	3.2e-05	3.2e-05	0.00891
GO:0042288	MHC class I protein binding	7	6.5e-05	6.5e-05	6.5e-05	0.18571
GO:0005516	calmodulin binding	5	8.7e-05	8.7e-05	8.7e-05	0.00091
GO:0046961	proton-transporting ATPase activity, rot...	16	0.00015	0.00015	0.00015	0.07025
GO:0005524	ATP binding	10	0.00020	0.00020	0.00020	0.70522
GO:0004971	AMPA glutamate receptor activity	12	0.00023	0.00023	0.00023	0.27219

(c) weight01

TERM.ID	Term	Level	classic	elim	weight01	parentchild
GO:0005215	transporter activity	3	1.6e-11	0.84654	1.00000	1.6e-11
GO:0008066	glutamate receptor activity	7	1.4e-08	0.08546	1.00000	4.3e-10
GO:0005515	protein binding	4	6.2e-08	0.01142	0.00102	4.6e-08
GO:0001882	nucleoside binding	5	3.4e-06	0.11574	0.03526	2.6e-07
GO:1901265	nucleoside phosphate binding	5	0.00010	0.73511	1.00000	4.4e-06
GO:0043168	anion binding	5	0.00026	0.97678	1.00000	1.4e-05
GO:0042625	ATPase activity, coupled to transmembran...	13	9.0e-12	1.00000	1.00000	4.1e-05
GO:0019829	cation-transporting ATPase activity	14	6.5e-12	0.02009	1.00000	6.1e-05
GO:0005198	structural molecule activity	3	0.00011	0.17612	0.48464	0.00011
GO:0097367	carbohydrate derivative binding	4	3.7e-05	0.58353	1.00000	0.00012

(d) parentchild

Table A5: ARC complexes GOMF enrichment result

Bibliography

[1] Russ B. Altman. Introduction to Translational Bioinformatics Collection. *PLoS Computational Biology*, 8(12), 2012.

[2] Atul J. Butte. Translational Bioinformatics: Coming of Age. *Journal of the American Medical Informatics Association*, 15(6):709–714, 2008.

[3] John D Osborne, Jared Flatow, Michelle Holko, Simon M Lin, Warren a Kibbe, Lihua Julie Zhu, Maria I Danila, Gang Feng, and Rex L Chisholm. Annotating the human genome with Disease Ontology. *BMC genomics*, 10 Suppl 1:S6, jan 2009.

[4] Paea LePendu, Mark A. Musen, and Nigam H. Shah. Enabling enrichment analysis with the Human Disease Ontology. *Journal of Biomedical Informatics*, 44(SUPPL. 1):S31–8, dec 2011.

[5] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008.

[6] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2009.

[7] Robert Plomin and Leonard C. Schalkwyk. Microarrays, 2007.

[8] WHO. *International Statistical Classification of Diseases and Related Health Problems (International Classification of Diseases)(ICD) 10th Revision - Version:2010*, volume 1. 2010.

[9] Margaret H Coletti and Howard L Bleich. Medical Subject Headings Used to Search the Biomedical Literature. *Journal of the American Medical Informatics Association*, 8(4):317–323, 2001.

[10] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270, 2004.

[11] GeneOntologyConsortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue):258D–261, 2004.

[12] Rachael P Huntley, Tony Sawford, Maria J Martin, and Claire O'Donovan. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *GigaScience*, 3(1):4, 2014.

[13] Lynette Hirschman, Gully A P C Burns, Martin Krallinger, Cecilia Arighi, K. Bretonnel Cohen, Alfonso Valencia, Cathy H. Wu, Andrew Chatr-Aryamontri, Karen G. Dowell, Eva Huala, Analia Lourenco, Robert Nash, Anne Lise Veuthey, Thomas Wiegers, and Andrew G. Winter. Text mining for the biocuration workflow. *Database*, 2012, 2012.

[14] K. G. Dowell, M. S. McAndrews-Hill, D. P. Hill, H. J. Drabkin, and J. A. Blake. Integrating text mining into the MGI biocuration workflow. *Database*, 2009, 2009.

[15] Pascale Gaudet, Michael S. Livstone, Suzanna E. Lewis, and Paul D. Thomas. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, 12(5):449–462, 2011.

[16] Da Wei Huang, Brad T. Sherman, and Richard a. Lempicki. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37:1–13, 2009.

[17] Barry Smith, Jennifer Williams, and Steffen Schulze-Kremer. The ontology of the gene ontology. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 609–13, 2003.

[18] Nicola Guarino. Formal Ontology and Information Systems. *Proceedings of the first international conference*, 46(June):3–15, 1998.

[19] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928, 1995.

[20] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[21] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25:1251–5, 2007.

[22] Mark A Musen, Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Christopher G Chute, Margaret-Anne Story, and Barry Smith. The National Center for Biomedical Ontology. *Journal of the American Medical Informatics Association : JAMIA*, 19:190–5, 2012.

[23] Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. The role of ontologies in biological and biomedical research: A functional perspective. *Briefings in Bioinformatics*, 16(6):1069–1080, 2015.

[24] Katherine Munn and Barry Smith. *Applied Ontology An Introduction*. 2009.

[25] Terrence F Meehan, Anna Masci, Amina Abdulla, Lindsay G Cowell, Judith A Blake, Christopher J Mungall, and Alexander D Diehl. Logical Development of the Cell Ontology. *BMC Bioinformatics*, 12(1):6, 2011.

[26] Grigoris Antoniou and Frank Van Harmelen. OWL Web Ontology Language. *Handbook on Ontologies in Information Systems*, 2007(September):157–160, 2004.

[27] Eric Antezana, Mikel Egaea, Bernard De Baets, Martin Kuiper, and Vladimir Mironov. ONTO-PERL: An API for supporting the development and analysis of bio-ontologies. *Bioinformatics*, 24(6):885–887, 2008.

[28] Junko Tanouo, Masatoshi Yoshikawa, and Shunsuke Uemura. The GeneAround GO viewer. *Bioinformatics*, 18(12):1705–1706, 2002.

[29] Guillaume Obozinski, Gert Lanckriet, Charles Grant, Michael I Jordan, and William Stafford Noble. Consistent probabilistic outputs for protein function prediction. *Genome biology*, 9 Suppl 1:S6, 2008.

[30] Babak Shahbaba and Radford M Neal. Gene function classification using Bayesian models with hierarchy-based priors. *BMC bioinformatics*, 7:448, 2006.

[31] R. Eisner, B. Poulin, D. Szafron, P. Lu, and R. Greiner. Improving Protein Function Prediction using the Hierarchical Structure of the Gene Ontology. *2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–10, 2005.

[32] Giorgio Valentini. True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):832–847, 2011.

[33] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22:1600–1607, 2006.

[34] The Gene Ontology Consortium. *GO Ontology Structure Guide*, 2016 (accessed September 4, 2016). http://geneontology.org/page/ontology-structure.

[35] Jane Lomax. Get ready to GO! A biologist's guide to the Gene Ontology. *Briefings in Bioinformatics*, 6(3):298–304, 2005.

[36] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E. Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[37] Nophar Geifman, Alon Monsonego, and Eitan Rubin. The Neural/Immune Gene Ontology: clipping the Gene Ontology for neurological and immunological systems. *BMC bioinformatics*, 11:458, January 2010.

[38] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael a Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102:15545–50, 2005.

[39] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40:D940–6, 2012.

[40] Sebastian Köhler, Sandra C. Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, Isabelle Bailleul-Forestier, Graeme C M Black, Danielle L. Brown, Michael Brudno, Jennifer Campbell, David R. Fitzpatrick, Janan T. Eppig, Andrew P. Jackson, Kathleen Freson, Marta Girdea, Ingo Helbig, Jane a. Hurst, Johanna Jähn, Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour, Christa L. Martin, Celia Moss, Andrew Mumford, Willem H. Ouwehand, Soo Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O M Wilkie, Caroline F. Wright, Anneke T. Vulto-Van Silfhout, Nicole De Leeuw, Bert B a De Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, and Peter N. Robinson. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42:966–974, 2014.

[41] Ashutosh Malhotra, Erfan Younesi, Michaela Gündel, Bernd Müller, Michael T Heneka, and Martin Hofmann-Apitius. ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 10:238–46, 2014.

[42] Purvesh Khatri and Sorin Drghici. Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*, 21:3587–3595, 2005.

[43] William K. McCoubrey, James F. Ewing, and Mahin D. Maines. Human heme oxygenase-2: Characterization and expression of a full-length cDNA and evidence suggesting that the two HO-2 transcripts may differ by choice of polyadenylation signal. *Archives of Biochemistry and Biophysics*, 295(1):13–20, 1992.

[44] William A Baumgartner, K Bretonnel Cohen, Lynne M Fox, George Acquaah-Mensah, and Lawrence Hunter. Manual curation is not sufficient for annotation

of genomic databases. *Bioinformatics (Oxford, England)*, 23(13):i41–i48, jul 2007.

[45] a. Bravo, J. Pinero, N. Queralt, M. Rautschka, and L. I. Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *bioRxiv*, page 007443, 2014.

[46] Noa Rappaport, Noam Nativ, Gil Stelzer, Michal Twik, Yaron Guan-Golan, Tsippi Iny Stein, Iris Bahir, Frida Belinky, C. Paul Morrey, Marilyn Safran, and Doron Lancet. MalaCards: An integrated compendium for diseases and their annotation. *Database*, 2013, 2013.

[47] Allan Peter Davis, Cynthia Grondin Murphy, Robin Johnson, Jean M. Lay, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L. King, Michael C. Rosenstein, Thomas C. Wiegers, and Carolyn J. Mattingly. The comparative toxicogenomics database: Update 2013. *Nucleic Acids Research*, 41(D1), 2013.

[48] Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. The genetic association database. *Nature genetics*, 36(5):431–432, 2004.

[49] Y Hakak, J R Walker, C Li, W H Wong, K L Davis, J D Buxbaum, V Haroutunian, and a a Fienberg. Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*, 98:4746–4751, 2001.

[50] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–36, 2010.

[51] Ramin Homayouni, Kevin Heinrich, Lai Wei, and Michael W. Berry. Gene clustering by Latent Semantic Indexing of MEDLINE abstracts. *Bioinformatics*, 21(1):104–115, 2005.

[52] National Center for Biomedical Ontology. *NCBO Annotator web interface*, 2016. https://bioportal.bioontology.org/annotator.

[53] Da Wei Huang, Brad T. Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, Robert Stephens, Michael W. Baseler, H. Clifford Lane, and Richard a. Lempicki. DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, 35:169–175, 2007.

[54] Sorin Draghici, Purvesh Khatri, Pratik Bhavsar, Abhik Shah, Stephen a. Krawetz, and Michael a. Tainsky. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Research*, 31:3775–3781, 2003.

[55] Barry R Zeeberg, Weimin Feng, Geoffrey Wang, May D Wang, Anthony T Fojo, Margot Sunshine, Sudarshan Narasimhan, David W Kane, William C Reinhold, Samir Lababidi, Kimberly J Bussey, Joseph Riss, J Carl Barrett, and John N Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome biology*, 4:R28, 2003.

[56] a Zien, R Küffner, R Zimmer, and T Lengauer. Analysis of gene expression data with pathway scores. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 8:407–17, 2000.

[57] K Mirnics, F A Middleton, A Marquez, D A Lewis, and P Levitt. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron*, 28:53–67, 2000.

[58] Mark Gerstein and Ronald Jansen. The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? *Current Opinion in Structural Biology*, 10:574–584, 2000.

[59] Paul Pavlidis, Darrin P Lewis, and William Stafford Noble. Exploring gene expression data with class scores. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 474–485, 2002.

[60] Da Wei Huang, Richard A Lempicki, and Brad T Sherman. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4:44–57, 2009.

[61] Qi Zheng and Xiu-Jie Wang. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic acids research*, 36:W358–63, 2008.

[62] Marco Masseroli, Dario Martucci, and Francesco Pinciroli. GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic acids research*, 32:W293–300, 2004.

[63] Daniel Gatti, William Barry, Andrew Nobel, Ivan Rusyn, and Fred Wright. Heading Down the Wrong Pathway: on the Influence of Correlation within Gene Sets. *BMC genomics*, 11:574, 2010.

[64] Tirtha Das and Ross Cagan. Drosophila as a novel therapeutic discovery tool for thyroid cancer. *Thyroid : official journal of the American Thyroid Association*, 20:689–695, 2010.

[65] Jelle J. Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, 23:980–987, 2007.

[66] Xing Qiu, Lev Klebanov, and Andrei Yakovlev. Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Statistical applications in genetics and molecular biology*, 4:Article34, 2005.

[67] Xin Lu and David L Perkins. Re-sampling strategy to improve the estimation of number of null hypotheses in FDR control under strong correlation structures. *BMC bioinformatics*, 8:157, 2007.

[68] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstrå le, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler, and Leif C Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34:267–273, 2003.

[69] P. Tamayo, G. Steinhardt, A. Liberzon, and J. P. Mesirov. The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, pages 1–22, 2012.

[70] Irina Dinu, John D Potter, Thomas Mueller, Qi Liu, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad S Famulski, Philip Halloran, and Yutaka Yasui. Improving gene set analysis of microarray data by SAM-GS. *BMC bioinformatics*, 8:242, 2007.

[71] Jung H Kim, Alla Karnovsky, Vasudeva Mahavisno, Terry Weymouth, Manjusha Pande, Dana C Dolinoy, Laura S Rozek, and Maureen a Sartor. LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types. *BMC Genomics*, 13:526, 2012.

[72] Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson, and Peter J Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, 10:161, 2009.

[73] Homin K Lee, William Braynen, Kiran Keshav, and Paul Pavlidis. ErmineJ: tool for functional analysis of gene expression data sets. *BMC bioinformatics*, 6:269, 2005.

[74] Marcel Smid and Lambert C J Dorssers. GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics (Oxford, England)*, 20:2618–25, 2004.

[75] Sang Mun Chi, Jin Kim, Seon Young Kim, and Dougu Nam. ADGO 2.0: Interpreting microarray data and list of genes using composite annotations. *Nucleic Acids Research*, 39:302–306, 2011.

[76] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, 2009.

[77] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–9, 2005.

[78] Dougu Nam, Sang-Bae Kim, Seon-Kyu Kim, Sungjin Yang, Seon-Young Kim, and In-Sun Chu. ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics (Oxford, England)*, 22:2249–53, 2006.

[79] Alexey V Antonov, Thorsten Schmidt, Yu Wang, and Hans W Mewes. ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic acids research*, 36:W347–51, 2008.

[80] Choong-hyun Sun, Min-sung Kim, Youngwoong Han, and Gwan-su Yi. COFECO: composite function annotation enriched by protein complex data. *Nucleic acids research*, 37:W350–5, 2009.

[81] Daniel Tabas-Madrid, Ruben Nogales-Cadenas, and Alberto Pascual-Montano. GeneCodis3: A non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Research*, 40:W478–W483, 2012.

[82] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges, 2012.

[83] Evelyn B Camon, Daniel G Barrell, Emily C Dimmer, Vivian Lee, Michele Magrane, John Maslen, David Binns, and Rolf Apweiler. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC bioinformatics*, 6 Suppl 1:S17, 2005.

[84] Haiying Wang, Francisco Azuaje, Olivier Bodenreider, and Joaquín Dopazo. Gene Expression Correlation and Gene Ontology-Based Similarity: An Assessment of Quantitative Relationships. *Proceedings of the ... IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2004:25–31, 2004.

[85] Oliver D. King, Rebecca E. Foulger, Selina S. Dwight, James V. White, and Frederick P. Roth. Predicting gene function from patterns of annotation. *Genome Research*, 13:896–904, 2003.

[86] Donna Maglott, Joanna S Amberger, and Ada Hamosh. Online Mendelian Inheritance in Man (OMIM): A Directory of Human Genes and Genetic Disorders. *Nucleic Acids Research*, pages 1–7, 2002.

[87] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27:29–34, 1999.

[88] Gary D. Bader, Doron Betel, and Christopher W V Hogue. BIND: The Biomolecular Interaction Network Database, 2003.

[89] Robert D Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, and Alex Bateman. The Pfam protein families database. *Nucleic acids research*, 38:D211–D222, 2010.

[90] Tim Beiß barth, Terence P. Speed, Tim Beissbarth, Terence P. Speed, Tim Beiß barth, and Terence P. Speed. GOstat: Find statistically overrepresented Gene Ontologies with a group of genes. *Bioinformatics*, 20:1464–1465, 2004.

[91] Bing Zhang, Denise Schmoyer, Stefan Kirov, and Jay Snoddy. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC bioinformatics*, 5:16, 2004.

[92] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18:71–103, 2003.

[93] Yoav Benjamini and Yosef Hochberg. Controling the false discorevy rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

[94] N H Shah and N V Fedoroff. CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics (Oxford, England)*, 20:1196–7, 2004.

[95] C. I. Castillo-Davis and D. L. Hartl. GeneMerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19:891–892, 2003.

[96] Gabriel F. Berriz, Oliver D. King, Barbara Bryant, Chris Sander, and Frederick P. Roth. Characterizing gene sets with FuncAssociate. *Bioinformatics*, 19:2502–2504, 2003.

[97] Fátima Al-Shahrour, Ramón Díaz-Uriarte, and Joaquín Dopazo. FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20:578–580, 2004.

[98] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[99] Yosef Hochberg. A Sharper Bonferroni Procedure for Multiple Tests of Significance, 1988.

[100] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[101] Bradley Efron and Robert Tibshirani. Using Specially Designed Exponential Families for Density Estimation. *The Annals of Statistics*, 24:2431–2461, 1996.

[102] B Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99:96–104, 2004.

[103] Per Broberg. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC bioinformatics*, 6:199, 2005.

[104] Mette Langaas, B.H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:555–572, 2005.

[105] Stan Pounds and Cheng Cheng. Improving false discovery rate estimation. *Bioinformatics (Oxford, England)*, 20:1737–45, 2004.

[106] Huey-miin Hsueh, James J Chen, and Ralph L Kodell. Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *Journal of biopharmaceutical statistics*, 13:675–89, 2003.

[107] Baolin Wu, Zhong Guan, and Hongyu Zhao. Parametric and nonparametric FDR estimation revisited. *Biometrics*, 62:735–744, 2006.

[108] J G Liao, Yong Lin, Zachariah E Selvanayagam, and Weichung Joe Shih. A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics (Oxford, England)*, 20:2694–701, 2004.

[109] Stan Pounds and Stephan W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19:1236–1242, 2003.

[110] J Aubert, A Bar-Hen, J J Daudin, and S Robin. Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC bioinformatics*, 5:125, 2004.

[111] Stefanie Scheid and Rainer Spang. twilight; a Bioconductor package for estimating the local false discovery rate. *Bioinformatics (Oxford, England)*, 21:2921–2, 2005.

[112] Stefanie Scheid and Rainer Spang. A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 1:98–108, 2004.

[113] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100:9440–9445, 2003.

[114] J D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 479–498, 2002.

[115] John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 66:187–205, 2004.

[116] Bradley Efron. Correlation and Large-Scale Simultaneous Significance Testing. *Journal of the American Statistical Association*, 102:93–103, 2007.

[117] Nicolai Meinshausen. False Discovery Control for Multiple Tests of Association Under General Dependence. *Scandinavian Journal of Statistics*, 33:227–237, 2006.

[118] Peter H Westfall. Resampling-based multiple testing. *North*, pages 1359–1364, 1989.

[119] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

[120] C Li and W Wang. DNA-Chip Analyzer (dChip). In *The Analysis of Gene Expression Data: Methods and Software*, pages 120–141. 2003.

[121] V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98:5116–21, 2001.

[122] Nils Blüthgen, Karsten Brand, Branka Cajavec, Maciej Swat, Hanspeter Herzel, and Dieter Beule. Biological profiling of gene groups utilizing Gene Ontology. *Genome informatics. International Conference on Genome Informatics*, 16:106–15, 2005.

[123] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research*, 39, 2011.

[124] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai Su L Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33, 2005.

[125] J. X. Li, H. R. Zheng, C. N. Ji, X. W. Fei, M. Zheng, Y. J. Gao, Y. Ren, S. H. Gu, Y. Xie, and Y. M. Mao. A novel splice variant of human XRN2 gene is mainly expressed in blood leukocyte. *DNA Sequence*, 16:143–146, 2005.

[126] Brad T Sherman, Da Wei Huang, Qina Tan, Yongjian Guo, Stephan Bour, David Liu, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard a Lempicki. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC bioinformatics*, 8:426, 2007.

[127] KJ Bussey, David Kane, and Margot Sunshine. MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome . . .* , 4:R27, 2003.

[128] Andreu Alibés, Patricio Yankilevich, Andrés Cañada, and Ramón Díaz-Uriarte. IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC bioinformatics*, 8:9, 2007.

[129] Da Wei Huang, Brad T Sherman, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard a Lempicki. DAVID gene ID conversion tool. *Bioinformation*, 2:428–430, 2008.

[130] Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. Data driven ontology evaluation. *Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 641–644, 2004.

[131] Kevin Bretonnel Cohen, Martha Palmer, and Lawrence Hunter. Nominalization and alternations in biomedical language. *PLoS ONE*, 3(9), 2008.

[132] Burr Settles. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

[133] J Gregory Caporaso, William A Baumgartner, David A Randolph, K Bretonnel Cohen, and Lawrence Hunter. Rapid pattern development for concept recognition systems: application to point mutations. *Journal of Bioinformatics and Computational Biology*, 5(6):1233–1259, 2007.

[134] Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC bioinformatics*, 9 Suppl 3:S3, 2008.

[135] Robert Leaman and Graciela Gonzalez. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing.*, 663:652–663, 2008.

[136] a R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21, 2001.

[137] Clement Jonquet, Nigam H Shah, and Mark A Musen. The open biomedical annotator. *Summit on translational bioinformatics*, 2009:56–60, 2009.

[138] Dietrich Rebholz-Schuhmann, Miguel Arregui, Sylvain Gaudan, Harald Kirsch, and Antonio Jimeno. Text processing through web services: Calling Whatizit. *Bioinformatics*, 24(2):296–298, 2008.

[139] Joshua C. Denny, Jeffrey D. Smithers, Randolph A. Miller, and Anderson Spickard. "Understanding" medical school curriculum content using KnowledgeMap. *Journal of the American Medical Informatics Association*, 10(4):351–362, 2003.

[140] Lawrence H. Reeve and Hyoil Han. CONANN: An online biomedical concept annotator. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4544 LNBI, pages 264–279, 2007.

[141] Qinghua Zou, Wesley W Chu, Craig Morioka, Gregory H Leazer, and Hooshang Kangarloo. IndexFinder: a method of extracting key concepts from clinical texts for indexing. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 763–7, 2003.

[142] David Hancock, Norman Morrison, Giles Velarde, and Dawn Field. Terminizer Assisting Mark-Up of Text Using Ontological Terms. *Nature Precedings*, pages 22–22, 2009.

[143] Martijn J Schuemie, Rob Jelier, and Jan A Kors. Peregrine: Lightweight gene name normalization by dictionary lookup. *Proc of the Second BioCreative Challenge Evaluation Workshop*, pages 131–133, 2007.

[144] Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K Bretonnel Cohen, Lawrence E Hunter, and Karin Verspoor. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC bioinformatics*, 15:59, 2014.

[145] Alan R Aronson. Metamap: Mapping text to the umls metathesaurus. *Bethesda MD NLM NIH DHHS*, pages 1–26, 2006.

[146] National Library of Medicine. *MetaMap Data File Builder*, 2016. https://metamap.nlm.nih.gov/DataFileBuilder.shtml.

[147] Michael Tanenblatt, Anni Coden, and Igor Sominsky. The ConceptMapper Approach to Named Entity Recognition. *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, pages 546–551, 2010.

[148] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, and Lawrence E Hunter. Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13:161, 2012.

[149] Nigam H Shah, Nipun Bhatia, Clement Jonquet, Daniel Rubin, Annie P Chiang, and Mark A Musen. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC bioinformatics*, 10 Suppl 9:S14, 2009.

[150] Samuel Alan Stewart, Maia Elizabeth Von Maltzahn, and S S Raza Abidi. Comparing Metamap to MGrep as a tool for mapping free text to formal medical lexicons. In *Proceedings of the 1st international workshop on knowledge extraction & consolidation from social-media in conjunction with the 11th international semantic web conference (ISWC 2012), Boston, USA*, pages 63–77, 2012.

[151] Nipun Bhatia, Nigam H Shah, Daniel L Rubin, Annie P Chiang, Mark a Musen, and Clement Jonquet. Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap. *AMIA Summit on Translational Bioinformatics*, 10 Suppl 9(9:S14):S14, 2009.

[152] Padmini Srinivasan and Bisharah Libbus. Mining MEDLINE for implicit links between dietary substances and diseases. In *Bioinformatics*, volume 20, 2004.

[153] Carolina Perez-Iratxeta, Matthias Wjst, Peer Bork, and Miguel A Andrade. G2D: a tool for mining genes associated with disease. *BMC genetics*, 6(1):45, 2005.

[154] T K Jenssen, a Laegreid, J Komorowski, and E Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature genetics*, 28(1):21–8, 2001.

[155] Dimitar Hristovski, Borut Peterlin, Joyce A. Mitchell, and Susanne M. Humphrey. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74(2-4):289–298, 2005.

[156] Marco Masseroli, Osvaldo Galati, and Francesco Pinciroli. GFINDer: Genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Research*, 33(SUPPL. 2), 2005.

[157] Marco Masseroli, Osvaldo Galati, Mauro Manzotti, Karina Gibert, and Francesco Pinciroli. Inherited disorder phenotypes: controlled annotation and statistical analysis for knowledge mining from gene lists. *BMC bioinformatics*, 6 Suppl 4:S18, 2005.

[158] Marilyn Safran, Irina Dalah, Justin Alexander, Naomi Rosen, Tsippi Iny Stein, Michael Shmoish, Noam Nativ, Iris Bahir, Tirza Doniger, Hagit Krug, Alexandra Sirota-Madi, Tsviya Olender, Yaron Golan, Gil Stelzer, Arye Harel, and Doron Lancet. GeneCards Version 3: the human gene integrator. *Database : the journal of biological databases and curation*, 2010:baq020, 2010.

[159] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe MayPendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1):D896–D901, 2017.

[160] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, 26(8):1112–1118, 2010.

[161] EBI. *Ontology Xref Service*, 2017 (accessed July 28, 2017). http://www.ebi.ac.uk/spot/oxo/.

[162] Kyong-Ha Lee, Yoon-Joon Lee, Hyunsik Choi, Yon Dohn Chung, and Bongki Moon. Parallel data processing with MapReduce. *ACM SIGMOD Record*, 40(4):11, 2012.

[163] John D Osborne, Simon Lin, and Warren A Kibbe. Other riffs on cooperation are already showing how well a wiki could work. *Nature*, 446:856, 2007.

[164] Iakes Ezkurdia, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L. Tress. Multiple evidence strands suggest that theremay be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, 23:5866–5878, 2014.

[165] a Johnson and C O'Donnell. An open access database of genome-wide association resutls. *BMC Med Gent*, 10:6, 2009.

[166] Lucia a Hindorff, Praveen Sethupathy, Heather a Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri a Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–7, 2009.

[167] Michael Lichten. Genomics: Thoroughly modern meiosis. *Nature*, 454:421–2, 2008.

[168] Yuan Chen, Fiona Cunningham, Daniel Rios, William M McLaren, James Smith, Bethan Pritchard, Giulietta M Spudich, Simon Brent, Eugene Kulesha, Pablo Marin-Garcia, Damian Smedley, Ewan Birney, and Paul Flicek. Ensembl variation resources. *BMC genomics*, 11:293, 2010.

[169] Arek Kasprzyk. BioMart: Driving a paradigm change in biological data management. *Database*, 2011:1–3, 2011.

[170] Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1), 2014.

[171] Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L Johnson, Christophe Roeder, Jinho D Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, William A Baumgartner, Michael Bada, Martha Palmer, and Lawrence E Hunter. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC bioinformatics*, 13(1):207, 2012.

[172] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: A knowledgebase of biological pathways. *Nucleic Acids Research*, 33, 2005.

[173] Huaiyu Mi, Anushya Muruganujan, and Paul D. Thomas. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41, 2013.

[174] Huaiyu Mi, Anushya Muruganujan, John T Casagrande, and Paul D Thomas. Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, 8:1551–66, 2013.

[175] Frances Balkwill. Tumour necrosis factor and cancer. *Nature reviews. Cancer*, 9:361–71, 2009.

[176] K J Tracey and A Cerami. Tumor necrosis factor, other cytokines and disease. *Annual review of cell biology*, 9:317–343, 1993.

[177] Jilong Yang, Da Yang, Yan Sun, Baocun Sun, Guowen Wang, Jonathan C. Trent, Dejka M. Araujo, Kexin Chen, and Wei Zhang. Genetic amplification of the vascular endothelial growth factor (VEGF) pathway genes, including VEGFA, in human osteosarcoma. *Cancer*, 117:4925–4938, 2011.

[178] Elazar Zelzer, Roni Mamluk, Napoleone Ferrara, Randall S Johnson, Ernestina Schipani, and Bjorn R Olsen. VEGFA is necessary for chondrocyte survival during bone development. *Development (Cambridge, England)*, 131:2161–2171, 2004.

[179] Sylvanie Surget, Marie P. Khoury, and Jean Christophe Bourdon. Uncovering the role of p53 splice variants in human malignancy: A clinical perspective, 2013.

[180] Y M Irshaid, M A Abujbara, K M Ajlouni, M El-Khateeb, and Y B Jarrar. N-acetyltransferase-2 genotypes among Jordanian patients with diabetes mellitus. *Int J Clin Pharmacol Ther*, 51(7):593–599, 2013.

[181] Bernard Friedenson. The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. *BMC cancer*, 7:152, 2007.

[182] Marlene Rabinovitch and Vera Dunlevie. Rescuing the BMPR2 Pathway: How and Where Can We Intervene?

[183] Hensin Tsao and Kristin Niendorf. Genetic testing in hereditary melanoma, 2004.

[184] D. T. Bishop. Geographical Variation in the Penetrance of CDKN2A Mutations for Melanoma. *CancerSpectrum Knowledge Environment*, 94(12):894–903, 2002.

[185] Irene Orlow, Colin B Begg, Javier Cotignola, Pampa Roy, Amanda J Hummer, Brian a Clas, Urvi Mujumdar, Rebecca Canchola, Bruce K Armstrong, Anne Kricker, Loraine D Marrett, Robert C Millikan, Stephen B Gruber, Hoda Anton-Culver, Roberto Zanetti, Richard P Gallagher, Terence Dwyer, Timothy R Rebbeck, Peter a Kanetsky, Homer Wilcox, Klaus Busam, Lynn From, and Marianne Berwick. CDKN2A germline mutations in individuals with cutaneous malignant melanoma. *The Journal of investigative dermatology*, 127(5):1234–1243, 2007.

[186] A. L. Borges, F. Cuellar, Puig-Butille J. A., M. Scarone, L. Delgado, C. Badenas, M. Mila, J. Malvehy, V. Barquet, J. Nunez, M. Laporte, G. Fernandez, P. Levrero, M. Martinez-Asuaga, and S. Puig. CDKN2A mutations in melanoma families from Uruguay. *British Journal of Dermatology*, 161(3):536–541, 2009.

[187] Robi Tacutu, Thomas Craig, Arie Budovsky, Daniel Wuttke, Gilad Lehmann, Dmitri Taranukha, Joana Costa, Vadim E Fraifeld, and João Pedro de Magalhães. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic acids research*, 41:D1027–33, 2013.

[188] Olivier Arnaiz, Agata Malinowska, Catherine Klotz, Linda Sperling, Michal Dadlez, France Koll, and Jean Cohen. Cildb: A knowledgebase for centrosomes and cilia. *Database*, 2009, 2009.

[189] Guy Divita, Tony Tse, and Laura Roth. Failure analysis of MetaMap transfer (MMTx). *Studies in Health Technology and Informatics*, 107:763–767, 2004.

[190] Joakim Nivre. Dependency Grammar and Dependency Parsing. Technical Report 1959, 2005.

[191] Roberto Navigli. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

[192] J. Pinero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015:bav028–bav028, 2015.

[193] Joyce A Mitchell, Alan R Aronson, James G Mork, Lillian C Folk, Susanne M Humphrey, and Janice M Ward. Gene indexing: Characterization and analysis of NLM's GeneRIFs. *AMIA Annual Symposium Proceedings*, pages 460–4, 2003.

[194] Esther P.M. Tjin, Richard W.J. Groen, Irma Vogelzang, Patrick W.B. Derksen, Melanie D. Klok, Helen P. Meijer, Susanne Van Eeden, Steven T. Pals, and Marcel Spaargaren. Functional analysis of HGF/MET signaling and aberrant HGF-activator expression in diffuse large B-cell lymphoma. *Blood*, 107(2):760–768, 2006.

[195] Christopher Slape, Leah Y. Liu, Sarah Beachy, and Peter D. Apian. Leukemic transformation in mice expressing a NUP98-HOXD13 transgene is accompanied by spontaneous mutations in Nras, Kras, and Cbl. *Blood*, 112(5):2017–2019, 2008.

[196] Ping He, Lyuba Varticovski, Elise D. Bowman, Junya Fukuoka, Judith A. Welsh, Koh Miura, Jin Jen, Edward Gabrielson, Elisabeth Brambilla, William D. Travis, and Curtis C. Harris. Identification of carboxypeptidase E and γ-glutamyl hydrolase as biomarkers for pulmonary neuroendocrine tumors by cDNA microarray. *Human Pathology*, 35(10):1196–1209, 2004.

[197] Cancer Research UK. *Cancer incidence by age*, 2017 (accessed July 28, 2017). http://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence.

[198] Manfred Fliegauf, Thomas Benzing, and Heymut Omran. When cilia go bad: cilia defects and ciliopathies. *Nature reviews. Molecular cell biology*, 8(11):880–93, 2007.

[199] Amjad Horani, Thomas W. Ferkol, Susan K. Dutcher, and Steven L. Brody. Genetics and biology of primary ciliary dyskinesia, 2015.

[200] Muqing Cao and Qing Zhong. Cilia in autophagy and cancer. *Cilia*, 5:4, feb 2015.

[201] Junmin Pan and William Snell. The primary cilium: keeper of the key to cell division. *Cell*, 129(7):1255–1257, 2007.

[202] Tamina Seeger-Nukpezah, Joy L. Little, Victoria Serzhanova, and Erica A. Golemis. Cilia and cilia-associated proteins in cancer, 2013.

[203] Jose L Badano, Norimasa Mitsuma, Phil L Beales, and Nicholas Katsanis. The ciliopathies: an emerging class of human genetic disorders. *Annual review of genomics and human genetics*, 7(May):125–148, 2006.

[204] Albert-lászló Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1):56–68, 2011.

[205] Albert-László Barabási. Network medicine–from obesity to the ”diseasome”. *The New England journal of medicine*, 357(4):404–407, 2007.

[206] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–8690, 2007.

[207] Vuk Janjic and Natasa Przulj. Biological function through network topology: A survey of the human diseasome. *Briefings in Functional Genomics*, 11(6):522–532, 2012.

[208] M Rautiainen, J Nuutinen, H Kiukaanniemi, and Y Collan. Ultrastructural changes in human nasal cilia caused by the common cold and recovery of ciliated epithelium. *The Annals of otology, rhinology, and laryngology*, 101(12):982–987, dec 1992.

[209] Sang Jun Han, Hee-Seong Jang, Jee In Kim, Joshua H Lipschutz, and Kwon Moo Park. Unilateral nephrectomy elongates primary cilia in the remaining kidney via reactive oxygen species. *Scientific Reports*, 6:22281, feb 2016.

[210] Jee In Kim, Jinu Kim, Hee-Seong Jang, Mi Ra Noh, Joshua H Lipschutz, and Kwon Moo Park. Reduction of oxidative stress during recovery accelerates normalization of primary cilia length that is altered after ischemic injury in murine kidneys. *American Journal of Physiology-Renal Physiology*, 304(10):F1283–F1294, 2013.

[211] Radek Cmejla, Jana Cmejlova, Helena Handrkova, Jiri Petrak, Kvetoslava Petrtylova, Vladimir Mihal, Jan Stary, Zdena Cerna, Yahia Jabali, and Dagmar Pospisilova. Identification of mutations in the ribosomal protein L5 (RPL5) and ribosomal protein L11 (RPL11) genes in Czech patients with Diamond-Blackfan anemia. *Hum Mutat*, 30(3):321–327, 2009.

[212] Anas M Alazami, Mohammed Zain Seidahmed, Fatema Alzahrani, Adam O Mohammed, and Fowzan S Alkuraya. Novel IFT122 mutation associated with impaired ciliogenesis and cranioectodermal dysplasia. *Molecular genetics & genomic medicine*, 2(2):103–6, 2014.

[213] M. Siedlinski, M. H. Cho, Per Bakke, A. Gulsvik, D. A. Lomas, W. Anderson, X. Kong, S. I. Rennard, Terri H. Beaty, J. E. Hokanson, J. D. Crapo, and E. K. Silverman. Genome-wide association study of smoking behaviours in patients with COPD. *Thorax*, 66(10):894–902, 2011.

[214] Corry-Anke Brandsma, Maarten van den Berge, Dirkje S Postma, Marnix R Jonker, Sharon Brouwer, Peter D Paré, Don D Sin, Yohan Bossé, Michel Laviolette, and Juha Karjalainen. A large lung gene expression study identifying fibulin-5 as a novel player in tissue repair in COPD. *Thorax*, pages thoraxjnl–2014, 2014.

[215] G. Deslee, J. C. Woods, C. M. Moore, L. Liu, S. H. Conradi, M. Milne, D. S. Gierada, J. Pierce, A. Patterson, R. A. Lewit, J. T. Battaile, M. J. Holtzman, J. C. Hogg, and R. A. Pierce. Elastin expression in very severe human COPD. *European Respiratory Journal*, 34(2):324–331, 2008.

[216] Cassandra M. Kelleher, Edwin K. Silverman, Thomas Broekelmann, Augusto A. Litonjua, Melvin Hernandez, Jody S. Sylvia, Joan Stoler, John J. Reilly, Harold A. Chapman, Frank E. Speizer, Scott T. Weiss, Robert P. Mecham, and Benjamin A. Raby. A functional mutation in the terminal exon of elastin in severe, early-onset chronic obstructive pulmonary disease. *American Journal of Respiratory Cell and Molecular Biology*, 33(4):355–362, 2005.

[217] W M Fitch. Distinguishing homologous from analogous proteins. *Systematic zoology*, 19:99–113, 1970.

[218] Toni Gabaldón and Eugene V. Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14:360–366, 2013.

[219] Ensembl Compara. *Protein trees and orthologies*, 2016 (accessed September 4, 2016). http://www.ensembl.org/info/genome/compara/homology_method.html.

[220] Astrid Jeibmann and Werner Paulus. Drosophila melanogaster as a model organism of brain diseases. *International Journal of Molecular Sciences*, 10:407–440, 2009.

[221] Mayuko Nishimura, Karen Ocorr, Rolf Bodmer, and Jérôme Cartry. Drosophila as a model to study cardiac aging. *Experimental gerontology*, 46:326–30, 2011.

[222] Aymeric Chartier, Béatrice Benoit, and Martine Simonelig. A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. *The EMBO journal*, 25:2253–2262, 2006.

[223] Matias Mosqueira, Gabriel Willmann, Hannele Ruohola-Baker, and Tejvir S. Khurana. Chronic hypoxia impairs muscle function in the drosophila model of duchenne's muscular dystrophy (DMD). *PLoS ONE*, 5:e13450, 2010.

[224] Morio Ueyama, Yoshihiro Akimoto, Tomomi Ichimiya, Ryu Ueda, Hayato Kawakami, Toshiro Aigaki, and Shoko Nishihara. Increased apoptosis of myoblasts in Drosophila model for the Walker-Warburg syndrome. *PloS one*, 5:e11557, 2010.

[225] Jennifer Grant, José W Saldanha, and Alex P Gould. A Drosophila model for primary coenzyme Q deficiency and dietary rescue in the developing nervous system. *Disease models & mechanisms*, 3:799–806, 2015.

[226] Ludovic Vial and Eric Déziel. The fruit fly as a meeting place for microbes. *Cell host & microbe*, 4:505–7, 2008.

[227] Isabella Vlisidou, Andrea J Dowling, Iwan R Evans, Nicholas Waterfield, Richard H Ffrench-Constant, and Will Wood. Drosophila embryos as model systems for monitoring bacterial infection in real time. *PLoS pathogens*, 5:e1000518, 2009.

[228] Yiorgos Apidianakis and Laurence G Rahme. Drosophila melanogaster as a model host for studying Pseudomonas aeruginosa infection. *Nature protocols*, 4:1285–1294, 2009.

[229] Renee D Read, Webster K Cavenee, Frank B Furnari, and John B Thomas. A drosophila model for EGFR-Ras and PI3K-dependent human glioma. *PLoS genetics*, 5:e1000374, 2009.

[230] Thomas Roeder, Kerstin Isermann, and Michael Kabesch. Drosophila in asthma research. *American Journal of Respiratory and Critical Care Medicine*, 179:979–983, 2009.

[231] Ronald P. Kühnlein. Drosophila as a lipotoxicity model organism - more than a promise? *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*, 1801:215–221, 2010.

[232] Keith D. Baker and Carl S. Thummel. Diabetic Larvae and Obese Flies-Emerging Studies of Metabolism in Drosophila, 2007.

[233] Aaron T Haselton and Yih-Woei C Fridell. Adult Drosophila melanogaster as a model for the study of glucose homeostasis. *Aging*, 2:523–6, 2010.

[234] Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx–relation extraction using dependency parse trees. *Bioinformatics (Oxford, England)*, 23(3):365–71, 2007.

[235] Robert Hoffmann and Alfonso Valencia. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21(SUPPL. 2), 2005.

[236] Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. *Journal of the American Medical Informatics Association*, 16(3):328–337, 2009.

[237] Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics*, 9:207, 2008.

[238] David McClosky, Sebastian Riedel, Mihai Surdeanu, Andrew McCallum, and Christopher D Manning. Combining joint models for biomedical event extraction. *BMC bioinformatics*, 13 Suppl 1(11):S9, 2012.

[239] Haibin Liu, Lawrence Hunter, Vlado Kešelj, and Karin Verspoor. Approximate Subgraph Matching-Based Literature Mining for Biomedical Events and Relations. *PLoS ONE*, 8(4), 2013.

[240] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 15:4–15, 2006.

[241] C Friedman, H Liu, L Shagina, S Johnson, and G Hripcsak. Evaluating the UMLS as a source of lexical knowledge for medical language processing. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 189–93, 2001.

[242] T C Rindflesch and M Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform*, 36(6):462–477, 2003.

[243] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of BioNLP'09 shared task on event extraction. *Proceedings of the Workshop on BioNLP Shared Task - BioNLP '09*, (June):1, 2009.

[244] Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, 78(12), 2009.

[245] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 4(Table 2):423–es, 2004.

[246] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K Kähäri, Stephen Keenan, Eugene Kulesha, Fergal J Martin, Thomas Maurel, William M McLaren, Daniel N Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J Trevanion, Alessandro Vullo, Steven P Wilder, Mark Wilson, Amonida Zadissa, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J P Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R Zerbino, and Stephen M J Searle. Variant Effect Predictor, 2014.

[247] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):R44, 2005.

[248] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[249] Pauline C. Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003.

[250] Ivan Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, (SUPPL.76), 2013.

[251] Paul D Thomas and Anish Kejariwal. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15398–403, 2004.

[252] Juyong Park, Deok-Sun Lee, Nicholas A Christakis, and Albert-László Barabási. The impact of cellular networks on disease comorbidity. *Molecular systems biology*, 5(262):262, 2009.

[253] Jesse Gillis and Paul Pavlidis. The role of indirect connections in gene networks in predicting function. *Bioinformatics*, 27(13):1860–1866, 2011.

[254] Jesse Gillis and Paul Pavlidis. The impact of multifunctional genes on guilt "by association "analysis. *PLoS ONE*, 6(2), 2011.

[255] Jesse Gillis and Paul Pavlidis. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3), 2012.

[256] Bevan Kai Sheng Chung and Dong-Yup Lee. Flux-sum analysis: a metabolite-centric approach for understanding the metabolic network. *BMC systems biology*, 3:117, 2009.

[257] Ming Lu, Qipeng Zhang, Min Deng, Jing Miao, Yanhong Guo, Wei Gao, and Qinghua Cui. An analysis of human microRNA and disease associations. *PLoS ONE*, 3(10), 2008.

[258] Seth Carbon, Amelia Ireland, Christopher J. Mungall, Shengqiang Shu, Brad Marshall, Suzanna Lewis, Jane Lomax, Chris Mungall, Benjamin Hitz, Rama Balakrishnan, Mary Dolan, Valerie Wood, Eurie Hong, and Pascale Gaudet. AmiGO: Online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009.

[259] Ashley J Waardenberg, Samuel D Basset, Romaric Bouveret, and Richard P Harvey. CompGO: an R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments. *BMC bioinformatics*, 16(1):275, 2015.

[260] Weijun Luo and Cory Brouwer. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, 2013.

[261] Guangchuang Yu, Li Gen Wang, Guang Rong Yan, and Qing Yu He. DOSE: An R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609, 2015.

[262] Emily Brookes, Inês De Santiago, Daniel Hebenstreit, Kelly J. Morris, Tom Carroll, Sheila Q. Xie, Julie K. Stock, Martin Heidemann, Dirk Eick, Naohito Nozaki, Hiroshi Kimura, Jiannis Ragoussis, Sarah A. Teichmann, and Ana Pombo. Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. *Cell Stem Cell*, 10:157–170, 2012.

[263] Zhixin Zhao, Wei Zhang, Bruce a Stanley, and Sarah M Assmann. Functional proteomics of Arabidopsis thaliana guard cells uncovers new stomatal signaling pathways. *The Plant cell*, 20:3210–3226, 2008.

[264] Sonja Reiland, Gaëlle Messerli, Katja Baerenfaller, Bertran Gerrits, Anne Endler, Jonas Grossmann, Wilhelm Gruissem, and Sacha Baginsky. Large-scale Arabidopsis phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks. *Plant physiology*, 150:889–903, 2009.

[265] Katja Baerenfaller, Jonas Grossmann, Monica a Grobei, Roger Hull, Matthias Hirsch-Hoffmann, Shaul Yalovsky, Philip Zimmermann, Ueli Grossniklaus, Wilhelm Gruissem, and Sacha Baginsky. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science (New York, N.Y.)*, 320:938–941, 2008.

[266] Adrian Alexa and Jorg Rahnenfuhrer. topGO: Enrichment analysis for Gene Ontology. *October*, 2010.

[267] Herve Pages, Marc Carlson, Seth Falcon, and Nianhua Li. *AnnotationDbi: Annotation Database Interface*, 2016. R package version 1.32.3.

[268] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *Nature reviews. Genetics*, 9(7):509–15, 2008.

[269] Steffen Grossmann, Sebastian Bauer, Peter N. Robinson, and Martin Vingron. Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. *Bioinformatics*, 23(22):3024–3031, 2007.

[270] Jun Ma, Maureen A Sartor, and H V Jagadish. Appearance frequency modulated gene set enrichment testing. *BMC bioinformatics*, 12:81, 2011.

[271] Jorg Rahnenfuhrer, Francisco S Domingues, Jochen Maydt, and Thomas Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical applications in genetics and molecular biology*, 3:Article16, 2004.

[272] Manway Liu, Arthur Liberzon, Won Kong Sek, Weil R. Lai, Peter J. Park, Isaac S. Kohane, and Simon Kasif. Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genetics*, 3(6):0958–0972, 2007.

[273] Jui-Hung Hung, Troy W Whitfield, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Identification of functional modules that correlate with phenotypic difference: the influence of network topology. *Genome biology*, 11(2):R23, 2010.

[274] W Link, U Konietzko, G Kauselmann, M Krug, B Schwanke, U Frey, and D Kuhl. Somatodendritic expression of an immediate early gene is regulated by synaptic activity. *Proceedings of the National Academy of Sciences of the United States of America*, 92(12):5734–8, 1995.

[275] D. E. Moga, M. E. Calhoun, A. Chowdhury, P. Worley, J. H. Morrison, and M. L. Shapiro. Activity-regulated cytoskeletal-associated protein is localized to recently activated excitatory synapses. *Neuroscience*, 125(1):7–11, 2004.

[276] H Husi, M a Ward, J S Choudhary, W P Blackstock, and S G Grant. Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. *Nature neuroscience*, 3(7):661–669, 2000.

[277] Oswald Steward and Paul F. Worley. Selective targeting of newly synthesized Arc mRNA to active synapses requires NMDA receptor activation. *Neuron*, 30(1):227–240, 2001.

[278] E Fernandez, M O Collins, R T Uren, M V Kopanitsa, N H Komiyama, M D Croning, L Zografos, J D Armstrong, J S Choudhary, and S G Grant. Targeted tandem affinity purification of PSD-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Molecular Systems Biology*, 5:269, 2009.

[279] Shaun M Purcell, Jennifer L Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm O'Dushlaine, Kimberly Chambert, Sarah E Bergen, Anna Kähler, Laramie Duncan, Eli Stahl, Giulio Genovese, Esperanza Fernández, Mark O Collins, Noboru H Komiyama, Jyoti S Choudhary, Patrik K E Magnusson, Eric Banks, Khalid Shakir, Kiran Garimella, Tim Fennell, Mark DePristo, Seth G N Grant, Stephen J Haggarty, Stacey Gabriel, Edward M Scolnick, Eric S Lander, Christina M Hultman, Patrick F Sullivan, Steven A McCarroll, and Pamela Sklar. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–90, 2014.

[280] Shoaib Chowdhury, Jason D. Shepherd, Hiroyuki Okuno, Gregory Lyford, Ronald S. Petralia, Niels Plath, Dietmar Kuhl, Richard L. Huganir, and Paul F. Worley. Arc/Arg3.1 Interacts with the Endocytic Machinery to Regulate AMPA Receptor Trafficking. *Neuron*, 52(3):445–459, 2006.

[281] Jason D. Shepherd, Gavin Rumbaugh, Jing Wu, Shoaib Chowdhury, Niels Plath, Dietmar Kuhl, Richard L. Huganir, and Paul F. Worley. Arc/Arg3.1 Mediates Homeostatic Synaptic Scaling of AMPA Receptors. *Neuron*, 52(3):475–484, 2006.

[282] Colin Mclean, He Xin, Ian T Simpson, and Douglas J Armstrong. Improved Functional Enrichment Analysis of Biological Networks using Scalable Modularity Based Clustering. *Journal of Proteomics & Bioinformatics*, 9(1):9–18, 2016.

[283] Andrew J Pocklington, Mark Cumiskey, J Douglas Armstrong, and Seth G N Grant. The proteomes of neurotransmitter receptor complexes form modular networks with distributed functionality underlying plasticity and behaviour. *Molecular systems biology*, 2:2006.0023, 2006.

[284] Alex Bayés, Louie N van de Lagemaat, Mark O Collins, Mike D R Croning, Ian R Whittle, Jyoti S Choudhary, and Seth G N Grant. Characterization of the proteome, diseases and evolution of the human postsynaptic density. SUPP. *Nature neuroscience*, 14(1):19–21, 2011.

[285] Edward L. Huttlin, Lily Ting, Raphael J. Bruckner, Fana Gebreab, Melanie P. Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, Rui Dong, Virginia Guarani, Laura Pontano Vaites, Alban Ordureau, Ramin Rad, Brian K. Erickson, Martin Wühr, Joel Chick, Bo Zhai, Deepak Kolippakkam, Julian Mintseris, Robert A. Obar, Tim Harris, Spyros Artavanis-Tsakonas, Mathew E. Sowa, Pietro De Camilli, Joao A. Paulo, J. Wade Harper, and Steven P. Gygi. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, 162(2):425–440, 2015.

[286] Stéphane Mélik-Parsadaniantz and William Rostène. Chemokines and neuromodulation. *Journal of Neuroimmunology*, 198(1-2):62–68, 2008.

[287] Khyber Saify and Mostafa Saadat. Non-random distribution of breast cancer susceptibility loci on human chromosomes., nov 2012.

[288] Mostafa Saadat. Chromosomal distribution of schizophrenia susceptibility loci. *Journal of Molecular Neuroscience*, 51(2):401–402, 2013.

[289] Asit B. Biswas and Frederick Furniss. Cognitive phenotype and psychiatric disorder in 22q11.2 deletion syndrome: A review. *Research in Developmental Disabilities*, 53-54:242–257, 2016.

[290] Anne S. Bassett and Eva W C Chow. Schizophrenia and 22q11.2 deletion syndrome, 2008.

[291] Caterina Mele, Giuseppe Remuzzi, and Marina Noris. Hemolytic uremic syndrome, 2014.

[292] Anna M Blom. Complement: Deficiency Diseases. *Encyclopedia of Life Sciences*, (800):1–8, 2010.

[293] Oana Mocan and Dan L Dumitracu. THE BROAD SPECTRUM OF CELIAC DISEASE AND GLUTEN SENSITIVE ENTEROPATHY. *Clujul Medical*, 89(3):335–342, 2016.

[294] Erik Thorsby and Benedicte A. Lie. HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms. In *Transplant Immunology*, volume 14, pages 175–182, 2005.

[295] Yogita Ghodke-Puranik and Timothy B. Niewold. Immunogenetics of systemic lupus erythematosus: A comprehensive review, 2015.

[296] Jozef Gécz, Cheryl Shoubridge, and Mark Corbett. The genetic landscape of intellectual disability arising from chromosome X, 2009.

[297] Patrick S Tarpey, Raffaella Smith, Erin Pleasance, Annabel Whibley, Sarah Edkins, Claire Hardy, Sarah O'Meara, Calli Latimer, Ed Dicks, Andrew Menzies, Phil Stephens, Matt Blow, Chris Greenman, Yali Xue, Chris Tyler-Smith, Deborah Thompson, Kristian Gray, Jenny Andrews, Syd Barthorpe, Gemma Buck, Jennifer Cole, Rebecca Dunmore, David Jones, Mark Maddison, Tatiana Mironenko, Rachel Turner, Kelly Turrell, Jennifer Varian, Sofie West, Sara Widaa, Paul Wray, Jon Teague, Adam Butler, Andrew Jenkinson, Mingming Jia, David Richardson, Rebecca Shepherd, Richard Wooster, M Isabel Tejada, Francisco Martinez, Gemma Carvill, Rene Goliath, Arjan P M de Brouwer, Hans van Bokhoven, Hilde Van Esch, Jamel Chelly, Martine Raynaud, Hans-Hilger Ropers, Fatima E Abidi, Anand K Srivastava, James Cox, Ying Luo, Uma Mallya, Jenny Moon, Josef Parnau, Shehla Mohammed, John L Tolmie, Cheryl Shoubridge, Mark Corbett, Alison Gardner, Eric Haan, Sinitdhorn Rujirabanjerd, Marie Shaw, Lucianne Vandeleur, Tod Fullston, Douglas F Easton,

Jackie Boyle, Michael Partington, Anna Hackett, Michael Field, Cindy Skinner, Roger E Stevenson, Martin Bobrow, Gillian Turner, Charles E Schwartz, Jozef Gecz, F Lucy Raymond, P Andrew Futreal, and Michael R Stratton. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nature genetics*, 41(5):535–43, 2009.

[298] Vincent Des Portes. X-linked mental deficiency. *Handbook of Clinical Neurology*, 111:297–306, 2013.

[299] Verneri Anttila, Dale R. Nyholt, Mikko Kallela, Ville Artto, Salli Vepsäläinen, Eveliina Jakkula, Annika Wennerström, Päivi Tikka-Kleemola, Mari A. Kaunisto, Eija Hämäläinen, Elisabeth Widén, Joseph Terwilliger, Kathleen Merikangas, Grant W. Montgomery, Nicholas G. Martin, Mark Daly, Jaakko Kaprio, Leena Peltonen, Markus Färkkilä, Maija Wessman, and Aarno Palotie. Consistently Replicating Locus Linked to Migraine on 10q22-q23. *American Journal of Human Genetics*, 82(5):1051–1063, 2008.

[300] Csilla Krausz, Antoni Riera Escamilla, and Chiara Chianese. Genetics of male infertility: From research to clinic, 2015.

[301] Jennifer F Hughes and David C Page. The Biology and Evolution of Mammalian Y Chromosomes. *Annual Review of Genetics*, 49(1):annurev–genet–112414–055311, 2015.

[302] Hiroyuki Yamagishi. The 22q11.2 deletion syndrome. *The Keio journal of medicine*, 51(2):77–88, 2002.

[303] Nicholas A. Bishop, Tao Lu, and Bruce A. Yankner. Neural mechanisms of ageing and cognitive decline. *Nature*, 464(7288):529–535, 2010.

[304] D Hanahan and R A Weinberg. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

[305] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation, 2011.

[306] Melvyn T. Chow, Andreas Möller, and Mark J. Smyth. Inflammation and immune surveillance in cancer, 2012.

[307] Glenn Dranoff. Cytokines in cancer pathogenesis and cancer therapy. *Nature Reviews Cancer*, 4(1):11–22, 2004.

[308] Drew Pardoll. Cancer Immunology. In *Abeloff's Clinical Oncology: Fifth Edition*, pages 78–97.e6. 2013.

[309] Gaia Novarino, Naiara Akizu, and Joseph G. Gleeson. Modeling human disease in humans: The ciliopathies, 2011.

[310] Rama Rao Damerla, George C. Gabriel, You Li, Nikolai T. Klena, Xiaoqin Liu, Yu Chen, Cheng Cui, Gregory J. Pazour, and Cecilia W. Lo. Role of cilia in structural birth defects: Insights from ciliopathy mutant mouse models. *Birth Defects Research Part C - Embryo Today: Reviews*, 102(2):115–125, 2014.

[311] S. Paige Taylor, Tiago J. Dantas, Ivan Duran, Sulin Wu, Ralph S. Lachman, Michael J. Bamshad, Jay Shendure, Deborah a. Nickerson, Stanley F. Nelson, Daniel H. Cohn, Richard B. Vallee, and Deborah Krakow. Mutations in DYNC2LI1 disrupt cilia function and cause short rib polydactyly syndrome. *Nature Communications*, 6:7092, 2015.

[312] Alex Reynolds Paz Polak, Rosa Karlic, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S. Lawrence and John A. Stamatoyannopoulos & Shamil R. Sunyaev Eric Rynes, Kristian Vlahovicek. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539):360–364, 2015.

[313] R J Youle and D P Narendra. Mechanisms of mitophagy. *Nat Rev Mol Cell Biol*, 12(1):9–14, 2011.

[314] KF Winklhofer and C Haass. Mitochondrial dysfunction in Parkinson's disease. *Biochimica et biophysica acta*, 1802(1):29–44, 2010.

[315] Joungil Choi, Avinash Ravipati, Vamshi Nimmagadda, Manfred Schubert, Rudolph J. Castellani, and James W. Russell. Potential roles of PINK1 for increased PGC-1α-mediated mitochondrial fatty acid oxidation and their associations with Alzheimer disease and diabetes. *Mitochondrion*, 18(1):41–48, 2014.

[316] L Samaranch, O Lorenzo-Betancor, J M Arbelo, I Ferrer, E Lorenzo, J Irigoyen, M A Pastor, C Marrero, C Isla, J Herrera-Henriquez, and P Pastor. PINK1-linked parkinsonism is associated with Lewy body pathology. *Brain*, 133(Pt 4):1128–1142, 2010.

[317] Henk J Blom, Gary M Shaw, Martin den Heijer, and Richard H Finnell. Neural tube defects and folate: case far from closed. *Nature reviews. Neuroscience*, 7(9):724–31, 2006.

[318] Roberto Bianco, Davide Melisi, Fortunato Ciardiello, and Giampaolo Tortora. Key cancer cell signal transduction pathways as therapeutic targets, 2006.

[319] Richard Sever and Joan S. Brugge. Signal transduction in cancer. *Cold Spring Harbor Perspectives in Medicine*, 5(4), 2015.

[320] Alexander Levitzki and Shoshana Klein. Signal transduction therapy of cancer, 2010.

[321] Edward a Sausville, Yusri Elsayed, Manish Monga, and George Kim. Signal transduction–directed cancer treatments. *Annual review of pharmacology and toxicology*, 43:199–231, 2003.

[322] Roi Avraham and Yosef Yarden. Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nature reviews. Molecular cell biology*, 12(2):104–17, 2011.

[323] V. P. Eswarakumar, I. Lax, and J. Schlessinger. Cellular signaling by fibroblast growth factor receptors, 2005.

[324] Alan R. Saltiel and C. Ronald Kahn. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*, 414(6865):799–806, 2001.

[325] Louis F Reichardt. Neurotrophin-regulated signalling pathways., 2006.

[326] Johanna Andrae, Radiosa Gallini, and Christer Betsholtz. Role of platelet-derived growth factors in physiology and medicine. *Genes & development*, 22(10):1276–312, 2008.

[327] Huaizeng Chen, Dafeng Ye, Xing Xie, Bingya Chen, and Weiguo Lu. VEGF, VEGFRs expressions and activated STATs in ovarian epithelial carcinoma. *Gynecologic Oncology*, 94(3):630–635, 2004.

[328] Claudia Wellbrock, Maria Karasarides, and Richard Marais. The RAF proteins take centre stage. *Nature reviews. Molecular cell biology*, 5(11):875–885, 2004.

[329] Brendan D. Manning and Lewis C. Cantley. AKT/PKB Signaling: Navigating Downstream, 2007.

[330] Randen L Patterson, Damian B van Rossum, Nikolas Nikolaidis, Donald L Gill, and Solomon H Snyder. Phospholipase C-gamma: diverse roles in receptor-mediated calcium signaling. *Trends in biochemical sciences*, 30(12):688–97, 2005.

[331] J W Rocco and D Sidransky. p16(MTS-1/CDKN2/INK4a) in cancer progression. *Experimental cell research*, 264(1):42–55, 2001.

[332] Kieran F. Harvey and Iswar K. Hariharan. The Hippo pathway. *Cold Spring Harbor Perspectives in Biology*, 4(8), 2012.

[333] K F Harvey, X Zhang, and D M Thomas. The Hippo pathway and human cancer. *Nat Rev Cancer*, 13(4):246–257, 2013.

[334] Jane I Lin, Carole L C Poon, and Kieran F Harvey. The Hippo size control pathway–ever expanding. *Sci Signal*, 6(259):pe4, 2013.

[335] Fa Xing Yu and Kun Liang Guan. The Hippo pathway: Regulators and regulations, 2013.

[336] Raphael Kopan. Notch signaling. *Cold Spring Harbor Perspectives in Biology*, 4(10), 2012.

[337] Ingrid Espinoza and Lucio Miele. Notch inhibitors for cancer treatment. *Pharmacology & therapeutics*, 139(2):95–110, 2013.

[338] Lisa a Catapano and Husseini K Manji. G protein-coupled receptors in major psychiatric disorders. *Biochimica et biophysica acta*, 1768(4):976–993, 2007.

[339] Tom H M Ottenhoff, Frank a W Verreck, Elgin G R Lichtenauer-Kaligis, Marieke a Hoeve, Ozden Sanal, and Jaap T van Dissel. Genetics, cytokines and human infectious disease: lessons from weakly pathogenic mycobacteria and salmonellae. *Nature genetics*, 32(1):97–105, 2002.

[340] S J van Deventer. Cytokine and cytokine receptor polymorphisms in infectious disease. *Intensive care medicine*, 26 Suppl 1:S98–S102, 2000.

[341] H Eric Xu. Family reunion of nuclear hormone receptors: structures, diseases, and drug discovery. *Acta Pharmacol Sin*, 36(1):1–2, jan 2015.

[342] Mh Eileen Tan, Jun Li, H Eric Xu, Karsten Melcher, and Eu-Leong Yong. Androgen receptor: structure, role in prostate cancer and drug discovery. *Acta pharmacologica Sinica*, 36(1):1–21, 2014.

[343] Kazuhiro Ikeda, Kuniko Horie-Inoue, and Satoshi Inoue. Identification of estrogen-responsive genes based on the DNA binding properties of estrogen receptors using high-throughput sequencing technology. *Acta pharmacologica Sinica*, 36(1):24–31, 2015.

[344] Mafei Xu, Jun Qin, Sophia Y Tsai, and Ming-jer Tsai. The role of the orphan nuclear receptor COUP-TFII in tumorigenesis. *Acta pharmacologica Sinica*, 36(1):32–6, 2015.

[345] Krista S Crider, Thomas P Yang, Robert J Berry, and Lynn B Bailey. Folate and DNA Methylation: A Review of Molecular Mechanisms and the Evidence for Folate's Role 1,2. *Adv. Nutr*, 3:21–38, 2012.

[346] Preeti Sharma and David M Kranz. Recent advances in T-cell engineering for use in immunotherapy. *F1000Research*, 5, 2016.

[347] C A Dinarello. Proinflammatory cytokines. *Chest*, 118(2):503–8, 2000.

[348] H F Chen, J Y Shew, H N Ho, W L Hsu, and Y S Yang. Expression of leukemia inhibitory factor and its receptor in preimplantation embryos. *Fertil Steril*, 72(4):713–719, 1999.

[349] Shigeru Saito. Cytokine cross-talk between mother and the embryo/placenta. In *Journal of Reproductive Immunology*, volume 52, pages 15–33, 2001.

[350] Andrew McGuire, James Brown, Carmel Malone, Ray McLaughlin, and Michael Kerin. Effects of Age on the Detection and Management of Breast Cancer. *Cancers*, 7(2):908–929, 2015.

[351] WHO. WHO — Breast cancer: prevention and control, 2016.

[352] Rintaro Hashizume, Mamoru Fukuda, Ichiro Maeda, Hiroyuki Nishikawa, Daisuke Oyake, Yukari Yabuki, Haruki Ogata, and Tomohiko Ohta. The RING Heterodimer BRCA1-BARD1 Is a Ubiquitin Ligase Inactivated by a Breast Cancer-derived Mutation. *Journal of Biological Chemistry*, 276(18):14537–14540, 2001.

[353] Maryou B Lambros, Rachael Natrajan, Felipe C Geyer, Maria a Lopez-Garcia, Konstantin J Dedes, Kay Savage, Magali Lacroix-Triki, Robin L Jones, Christopher J Lord, Spiros Linardopoulos, Alan Ashworth, and Jorge S Reis-Filho. PPM1D gene amplification and overexpression in breast cancer: a qRT-PCR and

chromogenic in situ hybridization study. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 23(10):1334–1345, 2010.

[354] Sheila Seal, Deborah Thompson, Anthony Renwick, Anna Elliott, Patrick Kelly, Rita Barfoot, Tasnim Chagtai, Hiran Jayatilake, Munaza Ahmed, Katarina Spanova, Bernard North, Lesley McGuffog, D Gareth Evans, Diana Eccles, Douglas F Easton, Michael R Stratton, and Nazneen Rahman. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nature genetics*, 38(11):1239–1241, 2006.

[355] Tina Thorslund, Michael J McIlwraith, Sarah a Compton, Sergey Lekomtsev, Mark Petronczki, Jack D Griffith, and Stephen C West. The breast cancer tumor suppressor BRCA2 promotes the specific targeting of RAD51 to single-stranded DNA. *Nature structural & molecular biology*, 17(10):1263–5, 2010.

[356] Katherine Stemke-Hale, Ana Maria Gonzalez-Angulo, Ana Lluch, Richard M. Neve, Wen Lin Kuo, Michael Davies, Mark Carey, Zhi Hu, Yinghui Guan, Aysegul Sahin, W. Fraser Symmans, Lajos Pusztai, Laura K. Nolden, Hugo Horlings, Katrien Berns, Mien Chie Hung, Marc J. Van De Vijver, Vicente Valero, Joe W. Gray, René Bernards, Gordon B. Mills, and Bryan T. Hennessy. An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer. *Cancer Research*, 68(15):6084–6091, 2008.

[357] Dan R Robinson, Yi-Mi Wu, Pankaj Vats, Fengyun Su, Robert J Lonigro, Xuhong Cao, Shanker Kalyana-Sundaram, Rui Wang, Yu Ning, Lynda Hodges, Amy Gursky, Javed Siddiqui, Scott a Tomlins, Sameek Roychowdhury, Kenneth J Pienta, Scott Y Kim, J Scott Roberts, James M Rae, Catherine H Van Poznak, Daniel F Hayes, Rashmi Chugh, Lakshmi P Kunju, Moshe Talpaz, Anne F Schott, and Arul M Chinnaiyan. Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nature genetics*, 45(12):1446–51, 2013.

[358] Susanne Taucher, Andreas Salat, Michael Gnant, Werner Kwasny, Brigitte Mlineritsch, Rainer Christian Menzel, Marianne Schmid, Michael G. Smola, Michael Stierer, Christoph Tausch, Arik Galid, Günther Steger, and Raimund Jakesz. Impact of pretreatment thrombocytosis on survival in primary breast cancer. *Thrombosis and Haemostasis*, 89(6):1098–1106, 2003.

[359] Ewa Sierko and Marek Z. Wojtukiewicz. Platelets and angiogenesis in malignancy, 2004.

[360] Inder Lal, Kim Dittus, and Chris E Holmes. Platelets, coagulation and fibrinolysis in breast cancer progression. *Breast cancer research : BCR*, 15(4):207, 2013.

[361] Hanna S Kuznetsov, Timothy Marsh, Beth A Markens, Zafira Castaño, April Greene-Colozzi, Samantha A Hay, Victoria E Brown, Andrea L Richardson,

Sabina Signoretti, Elisabeth M Battinelli, and Sandra S McAllister. Identification of Luminal Breast Cancers that Establish a Tumor Supportive Macroenvironment Defined by Pro-Angiogenic Platelets and Bone Marrow Derived Cells. *Cancer discovery*, 2(12):1150–1165, dec 2012.

[362] Theresa Placke, Melanie Örgel, Martin Schaller, Gundram Jung, Hans-Georg Rammensee, Hans-Georg Kopp, and Helmut Rainer Salih. Platelet-Derived MHC Class I Confers a Pseudonormal Phenotype to Cancer Cells That Subverts the Antitumor Reactivity of Natural Killer Immune Cells. *Cancer Research*, 72(2):440 LP – 448, jan 2012.

[363] Myriam Labelle, Shahinoor Begum, and RichardO Hynes. Direct Signaling between Platelets and Cancer Cells Induces an Epithelial-Mesenchymal-Like Transition and Promotes Metastasis. *Cancer Cell*, 20(5):576–590, 2011.

[364] Elisavet Paplomata and Ruth O'Regan. The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. *Therapeutic advances in medical oncology*, 6(4):154–66, 2014.

[365] Stephen M Edwards, Zsofia Kote-Jarai, Julia Meitz, Rifat Hamoudi, Questa Hope, Peter Osin, Rachel Jackson, Christine Southgate, Rashmi Singh, Alison Falconer, David P Dearnaley, Audrey Ardern-Jones, Annette Murkin, Anna Dowe, Jo Kelly, Sue Williams, Richard Oram, Margaret Stevens, Dawn M Teare, Bruce A J Ponder, Simon A Gayther, Doug F Easton, and Rosalind A Eeles. Two percent of men with early-onset prostate cancer harbor germline mutations in the BRCA2 gene. *American journal of human genetics*, 72(1):1–12, jan 2003.

[366] Elena Castro, Chee Goh, David Olmos, E. Saunders, Daniel Leongamornlert, Malgorzata Tymrakiewicz, Nadiya Mahmud, Tokhir Dadaev, Koveela Govindasami, Michelle Guy, Emma Sawyer, Rosemary Wilkinson, Audrey Ardern-Jones, Steve Ellis, Debra Frost, Susan Peock, D. Gareth Evans, Marc Tischkowitz, Trevor Cole, Rosemarie Davidson, Diana Eccles, Carole Brewer, Fiona Douglas, Mary E. Porteous, Alan Donaldson, Huw Dorkins, Louise Izatt, Jackie Cook, Shirley Hodgson, M. John Kennedy, Lucy E. Side, Jacqueline Eason, Alex Murray, Antonis C. Antoniou, Douglas F. Easton, Zsofia Kote-Jarai, and Rosalind Eeles. Germline BRCA mutations are associated with higher risk of nodal involvement, distant metastasis, and poor survival outcomes in prostate cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 31(14):1748–1757, 2013.

[367] J L Beebe-Dimmer, C Yee, M L Cote, N Petrucelli, N Palmer, C Bock, D Lane, I Agalliu, M L Stefanick, and M S Simon. Familial Clustering of Breast and Prostate Cancer and Risk of Postmenopausal Breast Cancer in the Women's Health Initiative Study. *Cancer*, 121(8):1265–1272, 2015.

[368] Jim van Os and Shitij Kapur. Schizophrenia. *Lancet*, 374(9690):635–45, 2009.

[369] P. Steullet, J.H. Cabungcal, A. Monin, D. Dwir, P. O'Donnell, M. Cuenod, and K.Q. Do. Redox dysregulation, neuroinflammation, and NMDA receptor hypofunction: A central hub in schizophrenia pathophysiology? *Schizophrenia Research*, 176(1):41–51, 2016.

[370] Guillermo Gonzalez-Burgos, Takanori Hashimoto, and David A. Lewis. Alterations of cortical GABA neurons and network oscillations in schizophrenia, 2010.

[371] Christine Konradi and Stephan Heckers. Molecular aspects of glutamate dysregulation: Implications for schizophrenia and its treatment, 2003.

[372] Sophie Erhardt, Lilly Schwieler, Linda Nilsson, Klas Linderholm, and Göran Engberg. The kynurenic acid hypothesis of schizophrenia. *Physiology and Behavior*, 92(1-2):203–209, 2007.

[373] Kenneth L. Davis, René S. Kahn, Grant Ko, and Michael Davidson. Dopamine in schizophrenia: A review and reconceptualization. *American Journal of Psychiatry*, 148(11):1474–1486, 1991.

[374] Alecia Willis, Hans Uli Bender, Gary Steel, and David Valle. PRODH variants and risk for schizophrenia, 2008.

[375] Joshua L. Roffman, Anthony P. Weiss, Thilo Deckersbach, Oliver Freudenreich, David C. Henderson, Donna H. Wong, Charles H. Halsted, and Donald C. Goff. Interactive effects of COMT Val108/158Met and MTHFR C677T on executive function in schizophrenia. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 147(6):990–995, 2008.

[376] A Kamiya, K Kubo, T Tomoda, M Takaki, R Youn, Y Ozeki, N Sawamura, U Park, C Kudo, M Okawa, C A Ross, M E Hatten, K Nakajima, and A Sawa. A schizophrenia-associated mutation of DISC1 perturbs cerebral cortex development. *Nat Cell Biol*, 7(12):1167–1178, 2005.

[377] Hae Jeong Park, Won Sub Kang, Joo-Ho Chung, and Jong Woo Kim. Association between promoter polymorphism (rs10789970) in 5-hydroxytryptamine receptor 3B and poor concentration in schizophrenia. *Psychiatry Research*, 219(1):235–237, sep 2014.

[378] Renan P Souza, Vincenzo de Luca, Herbert Y Meltzer, Jeffrey a Lieberman, and James L Kennedy. Influence of serotonin 3A and 3B receptor genes on clozapine treatment response in schizophrenia. *Pharmacogenetics and genomics*, 20(4):274–6, 2010.

[379] A. P. Rajkumar, B. Poonkuzhali, A. Kuruvilla, A. Srivastava, M. Jacob, and K. S. Jacob. Outcome definitions and clinical predictors influence pharmacogenetic associations between HTR3A gene polymorphisms and response to clozapine in patients with schizophrenia. *Psychopharmacology*, 224(3):441–449, 2012.

[380] Leonhard Lennertz, Michael Wagner, Ingo Frommann, Svenja Schulze-Rauschenbach, Anna Schuhmacher, Kai Uwe Kühn, Ralf Pukrop, Joachim Klosterkötter, Wolfgang Wölwer, Wolfgang Gaebel, Marcella Rietschel, Heinz Häfner, Wolfgang Maier, and Rainald Mössner. A coding variant of the novel serotonin receptor subunit 5-HT3E influences sustained attention in schizophrenia patients. *European Neuropsychopharmacology*, 20(6):414–420, 2010.

[381] Anna Schuhmacher, Rainald Mössner, Boris B Quednow, Kai-Uwe Kühn, Michael Wagner, Gabriela Cvetanovska, Dan Rujescu, Peter Zill, Hans-Jürgen Möller, Marcella Rietschel, Petra Franke, Wolfgang Wölwer, Wolfgang Gaebel, and Wolfgang Maier. Influence of 5-HT3 receptor subunit genes HTR3A, HTR3B, HTR3C, HTR3D and HTR3E on treatment response to antipsychotics in schizophrenia. *Pharmacogenetics and genomics*, 19(11):843–851, 2009.

[382] Khalid Choudhury, Andrew McQuillin, Vinay Puri, Jonathan Pimm, Susmita Datta, Srinivasa Thirumalai, Robert Krasucki, Jacob Lawrence, Nicholas J Bass, Digby Quested, Caroline Crombie, Gillian Fraser, Nicholas Walker, Haitham Nadeem, Sophie Johnson, David Curtis, David St Clair, and Hugh M D Gurling. A genetic association study of chromosome 11q22-24 in two different samples implicates the FXYD6 gene, encoding phosphohippolin, in susceptibility to schizophrenia. *American journal of human genetics*, 80(4):664–72, 2007.

[383] Adrienne C. Lahti, Martin A. Weiler, Tamara Michaelidis, Arti Parwani, and Carol A. Tamminga. Effects of ketamine in normal and schizophrenic volunteers. *Neuropsychopharmacology*, 25(4):455–467, 2001.

[384] Tina Hinton and Graham A R Johnston. The Role of GABAA Receptors in Schizophrenia. *Cellscience Reviews*, 5(1):180–194, 200 8.

[385] Kazu Nakazawa, Veronika Zsiros, Zhihong Jiang, Kazuhito Nakao, Stefan Kolata, Shuqin Zhang, and Juan E. Belforte. GABAergic interneuron origin of schizophrenia pathophysiology, 2012.

[386] Melis Inan, Timothy J. Petros, and Stewart A. Anderson. Losing your inhibition: Linking cortical GABAergic interneurons to schizophrenia, 2013.

[387] E Costa, J M Davis, E Dong, D R Grayson, A Guidotti, L Tremolizzo, and M Veldic. A GABAergic Cortical Deficit Dominates Schizophrenia Pathophysiology. *Critical Reviews in Neurobiology*, 16(1-2):1–23, 2004.

[388] Uwe Rudolph and Hanns Möhler. GABAA receptor subtypes: Therapeutic potential in Down syndrome, affective disorders, schizophrenia, and autism. *Annual review of pharmacology and toxicology*, 54(1):483–507, 2014.

[389] Ralf Brisch, Arthur Saniotis, Rainer Wolf, Hendrik Bielau, Hans-Gert Bernstein, Johann Steiner, Bernhard Bogerts, Katharina Braun, Zbigniew Jankowski, Jaliya Kumaratilake, Maciej Henneberg, and Tomasz Gos. The Role of Dopamine in Schizophrenia from a Neurobiological and Evolutionary Perspective: Old Fashioned, but Still in Vogue. *Frontiers in Psychiatry*, 5:47, may 2014.

[390] Francesca Amati, Michela Biancolella, Alessio Farcomeni, Stefania Giallonardi, Susana Bueno, Daniela Minella, Lucia Vecchione, Giovanni Chillemi, Alessandro Desideri, and Giuseppe Novelli. Dynamic changes in gene expression profiles of 22q11 and related orthologous genes during mouse development. *Gene*, 391(1-2):91–102, 2007.

[391] Rosanna Weksberg, Andrea C. Stachon, Jeremy A. Squire, Laura Moldovan, Jane Bayani, Stephen Meyn, Eva Chow, and Anne S. Bassett. Molecular characterization of deletion breakpoints in adults with 22q11 deletion syndrome. *Human Genetics*, 120(6):837–845, 2007.

[392] Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nature genetics*, 43(10):969–976, sep 2011.

[393] Juan C. Leza, Borja Bueno, Miquel Bioque, Celso Arango, Mara Parellada, Kim Do, Patricio O'Donnell, and Miguel Bernardo. Inflammation in schizophrenia: A question of balance, 2015.

[394] Roosmarijn C Drexhage, Esther M Knijff, Roos C Padmos, Leonie Van Der Heul-Nieuwenhuijzen, Wouter Beumer, Marjan a Versnel, and Hemmo a Drexhage. The mononuclear phagocyte system and its cytokine inflammatory networks in schizophrenia and bipolar disorder. *Expert review of neurotherapeutics*, 10(1):59–76, 2010.